

# SDSS Virtual Data Grid Challenge Problems: Cluster Finding, Correlation Functions and Weak Lensing

James Annis<sup>1</sup>, Steve Kent<sup>1</sup>, Alex Szalay<sup>2</sup>

<sup>1</sup>Experimental Astrophysics Group, Fermilab

<sup>2</sup>Department of Physics and Astronomy, The Johns Hopkins University

Draft v1.1 December 17, 2001

We present a roadmap to three SDSS data grid challenge problems. We present some considerations on astronomical virtual data, that the data are relatively complex, that intelligent choices of metadata can speed problem solving considerably, and that bookkeeping is important. The first challenge problem is finding clusters of galaxies in a galaxy catalog, a problem that presents a balanced compute and storage requirement. Second is the computation of correlation functions and power spectra of galaxy distributions, a problem that is compute intensive. Third is the calculation of weak lensing signals from the galaxy images, a problem that is storage intensive. Finally we examine the steps that would lead to grid enabled production of large astronomical surveys.

1	Introduction.....	2
1.1	The SDSS Data Sets .....	2
1.2	The Challenge Problems .....	3
1.3	Towards Grid Enabled Production.....	3
2	Astronomical Virtual Data.....	3
2.1	The Data Complexity .....	3
2.2	Metadata: Design for Speed.....	3
2.3	Derived Data: Design for Variations .....	4
3	The SDSS Challenge Problems .....	5
3.1	The SDSS Challenge Problem 1: Cluster Catalog Generation.....	5
3.2	The SDSS Challenge Problem 2: Spatial Correlation Functions and Power Spectra .....	6
3.3	The SDSS Challenge Problem 3: Weak Lensing .....	7
4	The SDSS Experience: Towards Grid Enabled Astronomical Surveys.....	7
4.1	SDSS Data Replication.....	8
4.2	The SDSS Pipeline Experience .....	8
4.2.1	Description of the Abstracted SDSS Pipeline.....	8
4.2.2	How To Proceed .....	9
4.2.3	Southern Coadd As A Possible Future Testbed .....	9
5	Conclusion .....	11

# 1 Introduction

The Sloan Digital Sky Survey is a project to map one quarter of the night sky (York et al. 2000). We are using a 150 Megapixel camera to obtain 5-bandpass images of the sky. We then target the 1 million brightest galaxies for spectroscopy, allowing us to produce 3-dimensional maps of the galaxies in the local universe. The final imaging mosaic will be 1 million by 1 million pixels.

This has required the construction of special purpose equipment. Astronomers based at the Apache Point Observatory near White Sands, New Mexico use a specially designed wide field 2.5m telescope to perform both the imaging and the spectroscopy, and a nearby 0.5m telescope to perform the imaging calibration. The camera and spectrographs are the largest in operation. The data reduction operation was designed from the start to handle a very large data set where every pixel is of interest.

We had to learn new techniques to acquire and reduce the data, borrowing from the experience in the high energy physics community. We now need to learn new techniques to analyze the large data sets, and expect to learn together with the GriPhyN collaboration.

## 1.1 The SDSS Data Sets

The SDSS, during an imaging night, produces data at a 8 Mbytes/s rate. Imaging nights occur perhaps 20-30 nights per year and spectroscopy occupies most of the rest of the nights not dominated by the full moon. Nonetheless, imaging data dominates the data sizes.

The fundamental SDSS data products are shown in table 1. The sizes are for the Northern Survey and the Southern Survey will roughly double the total amount of data.

**Table 1: The Data of the SDSS**

Data	Description	Data Size (Gigabytes)
Catalogs	Measured parameters of all objects	500
Atlas images	Cutouts about all detected objects	700
Binned sky	Sky after removal of detected objects	350
Masks	Regions of the sky not analyzed	350
Calibration	Calibration information	150
Frames	Complete corrected images	10,000

Our reduction produces complicated data. The catalog entry describing a galaxy has 120 members, including a radial profile. If a different measurement is needed, one can take the atlas images of the object and make a new measurement. If one desires to look for objects undetected by the normal processing, say low surface brightness galaxies, one can examine the binned sky. And one is always free to go back to the reduced image and try a different method of reduction.

The SDSS data sets are representative of the types of data astronomy produces and in particular the types that the NVO will face. We will be working closely with the NVO collaboration.

## 1.2 *The Challenge Problems*

The SDSS data allow a very wide array of analyses to be performed. Most involve the extraction of small data sets from the total SDSS data set. Some require the whole data, and some of these require computations beyond what is available from a SQL database. We have chosen three analyses to be SDSS challenge problems: these highlight the interesting domain problems of catalog versus pixel analyses, of high computation load, high storage load, and balanced analyses.

## 1.3 *Towards Grid Enabled Production*

The SDSS data were reduced using custom codes on large laboratory compute clusters. We can learn from the experience what is required to take the reduction into a grid reduction process, something that may be essential to handle the data streams expected from proposed surveys like the Large Synoptic Telescope, which is expected to take data at a rate of Terabytes per night.

# 2 **Astronomical Virtual Data**

## 2.1 *The Data Complexity*

Our data is relatively complicated, and we expect that this is a general feature of large surveys. We produce FITS files of the catalog data which contain 120 members including a radial profile (light measurements at a series of aperture sizes) and a large number of arrays (a measurement in each bandpass with an associated error). We produce atlas images, which are the pixel data cut out around detected galaxies, in all five band passes. These are optimal for re-measuring quantities on galaxies. We produce binned sky, which is binned copy of the pixel data for the images with all detected objects removed. The spectroscopy of the objects results in a catalog with another 100 members, and in 1-dimensional images of the spectra that are useful for measuring lines that the production pipeline does not attempt. Finally we have the files in two formats, the flat files in FITS format, and in a database format (two databases, currently).

## 2.2 *Metadata: Design for Speed*

The metadata requirements for SDSS catalog very naturally map from the concepts of the FastNPoint codes of Andrew Moore and collaborators (Moore et al.). In this world view, Grid Containers are not files or objects, but nodes of a kd-tree (or perhaps some other tree structure with better data insertion properties). In this view what matters for performance is the ability to know what is in the container without having to actually read the contents. Consider a metadata example listing the most useful quantities in astronomy:

Ra, Dec position on sky		bounding box
Z	redshift	bounding box
r	r-band brightness	bounding box
g-r	g-r color	bounding box
r-i	r-i color	bounding box

If the metadata includes the min, max, mean, standard deviation, and quartiles of the data, the execution time for a range search can be brought down from  $N^2$  to  $N \log N$ . The central ideas are exclusion (if the range to be searched for does not cross the bounding box, one need not read that container) and subsumption (if the range

to be searched for completely contains the bounding box, one needs the entire catalog, again not reading the container).

Furthermore, there are great possibilities for speed up if one is willing to accept an approximation or a given level of error. Clearly the majority of time in the range search above is spent going through the containers that have bounding boxes crossing the search; it is also true that often this affects the answer but little, as the statistics are dominated by the totally subsumed containers. Having the relevant metadata in principle allows the user to accept a level of error in return for speed.

Often what one is doing is to compare every object against every other object. The tree structure above gives considerable speed up; another comparable speedup is allowed if the objects of interest are themselves in containers with metadata allowing the exclusion and subsumption principles to operate.

These considerations also suggest that a useful derived dataset will be tree structures built against the Grid containers, with the relevant metadata built via time-consuming processing but then available quickly to later users.

### 2.3 *Derived Data: Design for Variations*

Most of astronomy is derived datasets. One of the clearest examples is the identification of clusters of galaxies. Nature has made a clean break at the scale of galaxies: galaxies and entities smaller are cleanly identifiable as single entities; above galaxies the entities are statistical. Given that, there are many different ways to identify clusters of galaxies. The SDSS currently is exploring 6 different cluster catalogs.

**Table 2: SDSS Cluster Finders**

Cluster Catalog	Description	Data
MaxBcg	Red luminous galaxies	Imaging catalog
Adaptive Matched Filter	Spatial/luminosity profiles	Imaging catalog
Voronoi	Voronoi tessellation	Imaging catalog
Cut and Enhance	Spatial/color search	Imaging catalog
C4	Color-color space	Imaging catalog
FOG	Velocity space overdensities	Spectroscopic catalog

Each of these catalogs are derived data sets. They may, in principle, be downloaded for existing regions, or the algorithm may be run at individual points in space, or a production run of the algorithm may be scheduled. It is worth pointing out that

- a. Each algorithms have changeable parameters,
- b. Each algorithm evolves and hence has version numbers,
- c. The underlying data can change as the reduction or calibration is re-performed.

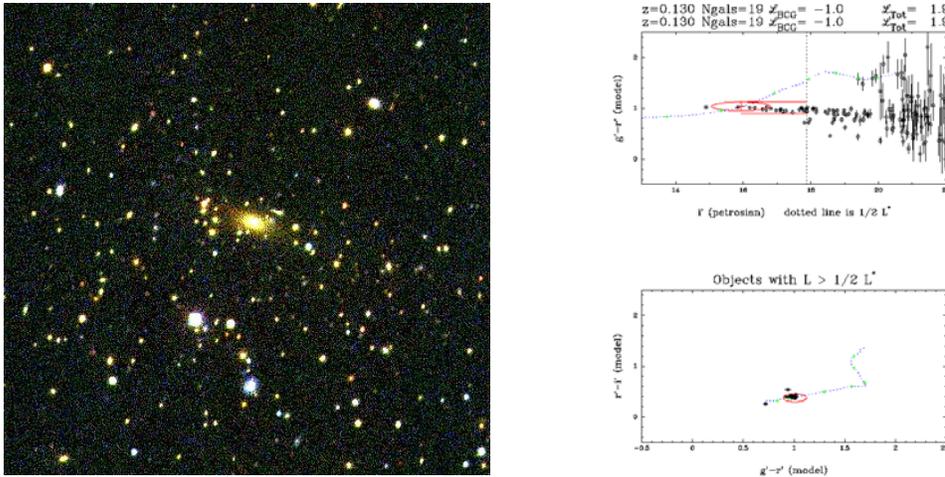
We thus point out that versioning and the associated bookkeeping is important.

Finally, we note that generically in astronomy one wishes to attach derived data to the underlying data. Here cluster finding is not a good example, and we will turn to the environment of individual galaxies. View the underlying galaxy data as a table; what astronomers generically wish to do is to add columns. Examples include counting red galaxies in some radius about each galaxy, counting all galaxies in some radius about each galaxy, summing the H-alpha emission from all galaxies with spectra in some radius about each galaxy, etc. The reigning technology in the SDSS is tables with row to row matching by position.

### 3 The SDSS Challenge Problems

#### 3.1 The SDSS Challenge Problem 1: Cluster Catalog Generation

The identification of clusters of galaxies in the SDSS galaxy catalog is a good example of derived data, and is naturally extendable to the idea of virtual data. We have chosen this as our first challenge problem.



**Figure 1:** A cluster of galaxies seen in a true color image on the left, and as a color-magnitude and color-color plot on the right. The plots on the right illustrate one cluster finding technique, a matched filter on the E/S0 ridgeline in color-luminosity space: the number of galaxies inside the red is the signal.

Clusters of galaxies are the largest bound structures in the universe; a good analogy is a hot gas cloud, where the molecules are galaxies. By counting clusters at a variety of redshifts as a function of mass, one is able to probe the evolution of structure in the universe. The number of the most massive clusters is a sensitive measure of the mass density  $\Omega_m$ ; combined with the cosmic microwave background measurements of the shape of the universe, these become a probe of the dark energy.

The basic procedure to find clusters is to count the number of galaxies within some range about a given galaxy. This is an  $N^2$  process, though with the use of metadata stored on trees it can be brought down to a  $N \log(N)$  problem. Note that the procedure is done for each galaxy in the catalog.

The problem is computationally expensive, though balanced with the I/O requirements; with the appropriate choices of parameters it can be made either an I/O bound problem or a CPU bound problem. The problem faces moderate storage problems: a hundred square degrees of SDSS data masses to 25 Gig. The problem can be made embarrassingly parallel as there is an outer bound to the apparent size of clusters of interest. The work proceeds through many stages and through many intermediate files that can be used as a form of checkpoint.

The problem is a good choice for the initial challenge problem as

1. cluster catalogs are a good example of derived data,

2. cluster catalog creation is roughly compute and storage balanced, and
3. it can be solved in interesting times on existing testbeds.

In terms of using GriPhyN tools, it exercises

1. the use of derived data catalogs and metadata,
2. replica catalogs,
3. transformation catalogs including DAG creation, and
4. to use existing cluster finding code advances in code migration must be made.

### 3.2 The SDSS Challenge Problem 2: Spatial Correlation Functions and Power Spectra

Our second challenge problem is aimed at computationally expensive measurements on catalog level data. We choose measurements on the spatial distribution of galaxies, which contain interesting cosmological information.

The correlation function of the positions of galaxies projected on the sky forms a Fourier transform pair with the spatial power spectrum. The power spectrum itself is of great interest in so much as the light from stars in galaxies traces the underlying mass both of normal matter and of the dark matter. If light traces mass, then when one measures the power spectra of galaxies one is measuring the power spectra of mass in the universe. The power spectrum may be predicted theoretically given a cosmology and mass and energy content of the universe, and thus this measurement explores very interesting quantities. The main uncertainty here is that it is known that the distribution of galaxies is biased away from the distribution of mass; exactly how much is a matter of some debate. The SDSS will allow these correlation functions to be measured and analyzed as a function of galaxy properties (e.g. magnitude, surface brightness, spectral type).

If the redshift of the objects are known, either from spectroscopy or by photometric redshift techniques, one is able to compute the power spectrum directly. This often involves an expensive SVD matrix calculation. The same push to measure the power spectrum as a function of galaxy properties exists, for the same reason; the relation of the galaxies to the underlying mass is uncertain.

The essential procedure is to count the distance from each galaxy to every other galaxy, accumulating the statistics. This is an  $N^2$  process (and higher order correlations equivalently higher order) though again metadata employing tree structures can cut the expense down to  $N \log(N)$ . The SVD matrix calculation is of order  $N^3$ . Neither program is particularly parallelizable, only made embarrassingly parallel by placing an arbitrary large angle cutoff. For the correlation function there is a further expensive step, that of the extensive randomization procedure that must be carried out in order to establish the statistical significance of the measurements.

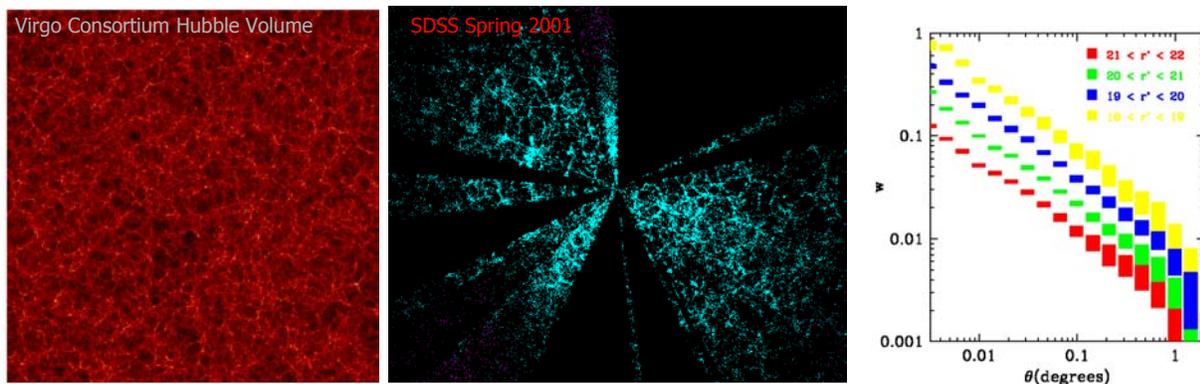


Figure 2: The correlation function. On the left panel, a numerical simulation of a Gigaparsec scale. In the middle, the SDSS redshift survey over a similar scale. On the right, the correlation function from the angular distribution of galaxies. The distribution of galaxies on the sky and in redshift contains cosmological information.

In both the correlation function measurement and the power spectrum measurement, the computation burden dominates the storage burden.

The correlation function challenge problem is an example of a catalog level analysis requiring large amounts of compute time. Each correlation function is computationally expensive, and very often the results are used as input to another layer of sophisticated codes. Correlation functions are an example of virtual data where a premium will be placed on locating existing materializations before requesting an expensive new materialization. If none can be found, then one enters the realm of resource discovery and job management. Locating existing materializations that are relevant is of course a very interesting problem: how does one construct a GriPhyN Google?

### ***3.3 The SDSS Challenge Problem 3: Weak Lensing***

Our third challenge problem is aimed at the storage bound problem of making new measurements on the pixel level data. We choose the problem of making optimal moments measurements of galaxies in the atlas images, which can be used to make weak lensing measurements of mass at cosmological distances.

Weak lensing is caused by mass distorting the path of light from background objects; the objects tend to align in concentric circles. Tend is the operative phrase; weak lensing is an inherently statistical program. One use for weak lensing is to measure the masses of the clusters. In the North one averages many clusters together to get signal. In the SDSS South, which is made considerably deeper by coadding many images together, individual clusters may have their mass measured. Two other measurements are equally interesting: measurements of the mass of the average galaxy, binned by interesting quantities, and measurements of the power spectrum of large scale structure induced weak lensing fluctuations.

In order to do weak lensing, one must make suitably weighted second moments analysis of each image. Most of the algorithmic magic lies in the exact weighting, some in the details of the moment analysis. This is an  $N^2$  process in the number of pixels, which of course is much larger than the number of galaxies. Despite this, the problem is weighted towards storage. The vast bulk of the atlas images that must be shepherded about dominates the problem. An analysis of the problem must include whether bandwidth is limited; if it is, compute power local to the data is preferred, if not, then the access to the distributed computing would be preferred.

The weak lensing challenge problem is a pixel level analysis that requires the moving of large amounts of data. Work will need to be done on data discovery, data and resource management.

## **4 The SDSS Experience: Towards Grid Enabled Astronomical Surveys**

In the long term, it is of interest to use GriPhyN tools in full scale production. The following two sections describe issues in production that map naturally onto Grid technologies.

## 4.1 SDSS Data Replication

Several sites wish to have local copies of the SDSS. Mostly this wish extends only to the catalog data, which will mass about a Terabyte at the end of the survey. We can expect that soon astronomers will wish to mirror the atlas images or even the reduced images.

The current model for transferring the catalogs involves sending tapes and sending disks. This is not without problems. The data live in the Objectivity based database SX, and the model has been to send loading files that the remote sites can load into their own SX database.

A different model would be to employ GDMP, which is nearly ideal for the purpose. It implements a subscription model that is well matched to the problem. It uses an efficient FTP, important when faced with 100 Gig scale transfers. It has transfer restart capability, very important when faced with the small pipe to Japan. The main problem here will be to convince our colleagues to spend the energy to bring up Globus and GDMP. Unless it is very easy to install and run, astronomers will return to the existing, if non-optimal, tools.

## 4.2 The SDSS Pipeline Experience

Data processing in SDSS is procedure oriented: start with raw data, run multi-stage processing programs, and save output files.

The data processing is file oriented. While we use the object-oriented database SX, the objects are used only internally. The natural unit of data for most SDSS astronomers is the “field”, one image of the sky, whose corresponding catalog has roughly 500 objects in it.

### 4.2.1 Description of the Abstracted SDSS Pipeline

The SDSS factory itself lives in the DP scripts that join the pipelines. There are 3 generic stages of the factory at each and every pipeline:

#### INPUTS:

1. A plan file - defining which set of data are to be processed.
2. Parameter files - tuning parameters applicable to this particular set of data.
3. Input files - that are the products of upstream pipelines.
4. Environment variables - defining which versions of pipelines are being run.

#### COMPUTATION:

1. Prep
  - a. generate plan file containing, for example, root directories for input and output
  - b. locate space
  - c. make relevant directories
  - d. stage the input data
  - e. make relevant sym links
  - f. register the resources reserved in a flat file database
  - g. Call submit
2. Submit
  - a. generate a shell script the fires off the batch system submit
  - b. Call ender
3. Ender

- a. Run a status verify job that checks if the submitted job did complete. The existence of the files that should have been created is necessary and almost sufficient for the next step to succeed.
- b. Run a pipeline verify job to generate QC information (e.g., the number of galaxies/image) that are given a sanity check. If success, call Prep for the follow-on pipeline.
- c. Run a "scrub" job that removes most files once they have been archived. The archiving is itself a "pipeline".

#### OUTPUTS:

1. Output files - the data products themselves.
2. Log and error files - log and error files
3. Quality control files - identify outputs so fatally flawed that subsequent pipelines cannot run.
4. A single status flag – 0, proceed to next pipeline, 1, hand intervention required.

These are daisy chained: the first invocation of Prep takes as an argument how many of the following pipelines to run in series.

### 4.2.2 How To Proceed

We note that the abstracted SDSS pipeline maps very well onto the DAG concepts of Condor. The path to the future lies in understanding the value added of using the DAG technology.

### 4.2.3 Southern Coadd As A Possible Future Testbed

The SDSS Southern Survey will in the end coadd various imaging runs on the same piece of sky to produce deep images. These images are suitable for making weak lensing measurements on individual clusters, as well as making more detailed weak lensing measurements on galaxies and large scale structure.

If one chose the problem of building weak lensing maps, one would exercise all of the pipeline structure outlined above. One vision of the SDSS/NVO team is to spin the SDSS data on a compute cluster and allow for demand driven re-reduction with the ever-improving versions of Photo. One cannot just run Photo, as it is the center of a long processing chain, but it is exactly this processing chain that must be made demand driven to solve the weak lensing map problem.

#### 4.2.3.1 Data Sizes

The components of any given 200 sq-degree coadd of the south come to:

catalogs	measured parameters of all objects	10 Gig
atlas images	cutouts around all objects	15 Gig
binned sky	sky leftover after cutouts, binned 4x4	7 Gig
masks	regions of the sky not analyzed	7 Gig
calibration	a variety of calibration information	3 Gig
frames	corrected images	Nx200 Gig

### 4.2.3.2 Design Detail

1. given an area on which to coadd,
  - a. Take user input
2. find relevant reduced data
  - a. Query either a db or a flat file for the the runs that have been observed and extract the list of runs that are of the right strip.
  - b. Using that list, determine which runs overlap the given area. Involves the calculation of a spatial intersection.
  - c. Apply cuts against data that are not of high enough quality to use (even though other parts of the run may be.)
  - d. Estimate how much data is involved.  
Metadata: a list of runs and portions of runs involved
3. find disk space and compute power for input data, processing, and outputs
  - a. Query the local network for available machines
  - b. Query the local network for available storage
  - c. Reserve the resources for a period of time  
Metadata: a list of machines
4. extract relevant reduced data from storage
  - a. From some knowledge base, determine for each run if the data is:
    - 1) on archival disk
    - 2) on production disk
    - 3) on fast tape
    - 4) on slow tape
  - b. Arrange to get the data off the media and onto the production machines  
Metadata: a list of portions of runs and where they live on production disk
5. build mapping function from calibration data
  - a. An image is a 1361x2048 array of pixels. In general two images of the "same" piece of sky will be:
    - 1) slightly offset in both x and y
    - 2) slightly rotated
    - 3) have different small distortions as a function of x,y superimposed
    - 4) have different conversion between pixel value and energy flux
  - b. These translations, rotations, distortions, and scalings are calculateable from a set of calibration information that may or may not be kept as a header to the data itself.
  - c. Find the calibrations, and build the mapping function for each pixel  
Metadata: computed mapping functions to be applied to the data
6. perform coadd to create new reduced data
  - a. Load a set of co-located images into memory
  - b. Apply mapping function
  - c. Perform median or average on stack of pixels (in truth: apply very clever algorithm to make maximal use of information and minimize noise)

- d. Save output to disk.  
Metadata: Location of output data
- 7. run full SDSS pipeline on new data set
  - a. arrange for all necessary input files to be in place
  - b. arrange for all 10 pipelines to run
  - c. watch success or failure of the pipelines
  - d. save resulting outputs to disk
 Metadata: Location of output data
- 8. keep track of intermediate coadd catalogs and data
  - a. The output of 6. is to be preserved, as 7. can be done multiple times
  - b. The output of 7. is to be preserved
  - c. Each time new data comes in, the above is done.
- 9) Run weak lensing analysis on the intermediate coadded south
  - a) locate atlas images on disk
  - b) compute optimal shape parameter
  - c) produce shape catalog

## 5 Conclusion

We have outlined the data of the SDSS, and how we might approach it with GriPhyN tools. Starting with cluster finding as an initial virtual data problem, we move through correlation function and power spectrum problems which push computational and virtual data recovery issues, to a weak lensing analysis on the pixels which push data management issues. We then lay out possible avenues for work leading to the incorporation of Grid tools in survey production.