Draft version October 19, 2025
Typeset using LaTeX **preprint** style in AASTeX631

# The New Age of Big Data In Astronomy: A Review of the SKA & Rubin

Mathew Icho
*The University of Illinois at Urbana-Champaign*

## ABSTRACT

I'm making my abstract my to do list for now

**1. Fix the SDSS part based on provided notes, also make real-time processing section**

**2. Find a paper or part of it that describes how SDSS stores the data. I don't think I've dont that sufficiently**

**3. Do the LSST part**

**4. Do the Results section, look at data collected. I plan on coding A LOT for this section**

**5. Redo the Du Pont section, it's outdated**

## Contents

## 1. INTRODUCTION

The concept of data has long been central throughout the history of astronomy. Data allows scientists to discover natural laws in the universe, have control over events, and make reliable predictions. It has played a critical role in other time-sensitive fields such as medicine and engineering, where accurate data is essential for decision-making and design. Although the nature of data varies fundamentally across different fields, one trend has remained consistent: the continual evolution of data science. As explained in The Fourth Paradigm (**?**), this evolution can be characterized through four successive paradigms. In the following sections, I describe the progression of data acquisition across these paradigms and illustrate each using examples from astronomy. I will then explain how SKA and LSST fit into this trajectory and exemplify the emerging era of data-intensive discovery.

### 1.1. *The Paradigms of Data Science*

The first and most primitive paradigm, as described by (**?**), is empirical evidence. Empirical evidence refers to data collected through traditional means, such as direct observation or experimentation. The primary purpose of empirical evidence is to identify patterns that allow scientists to develop a fundamental understanding of the natural world. Throughout much of human history, empirical evidence has dominated knowledge generation. An example of the first paradigm in astronomy is the career of Tycho Brahe, a Danish astronomer. Throughout his career in the 16th century, Brahe collected and cataloged data on the position of astronomical bodies using naked-eye observations. Tycho Brahe's catalogue was accurate to only around 1' precision and took decades to acquire (**?**). However, empirical evidence can be compromised by human error, the precision of

the instruments, and, most importantly, the relatively slow pace of data acquisition compared to subsequent paradigms.

The second paradigm is analytical evidence. Analytical evidence is obtained by constructing mathematical formulas and theoretical frameworks based on empirical data (**?**). Unlike the empirical evidence, which merely demonstrates that phenomena occur, the second paradigm seeks to explain why they occur. An example of the second paradigm in astronomy is the work of Johannes Kepler, who used Brahe's empirical observations to derive the laws of planetary motion (**?**). By transforming raw observational data into mathematical laws, Kepler exemplified how analytical evidence advances scientific understanding beyond description to explanation.

The third paradigm is simulation evidence (**?**), a relatively recent development. Simulation models natural phenomena that are too complex to model analytically or compute by hand. It allows interpolation and extrapolation of data using computational techniques grounded in known physical laws. For example, in astronomy, N-body simulations are used to study the complex dynamical evolution of planetary systems and galaxies.

The fourth and most recent paradigm is data-intensive science (**?**). This paradigm is characterized by the unprecedented scale, velocity, and complexity of data acquisition, driven in part by exponential advances in computational power and detector technologies, often associated with Moore's law (**?**). Unlike earlier paradigms, which focused on observation, theory, or simulation, data-intensive science emphasizes the ability to manage, analyze, and interpret vast datasets that exceed the capacity of traditional methods. While this exponential growth in data has enabled transformative discoveries, it also introduces significant challenges related to storage, processing, and accessibility.

### 1.2. *The Rise of Big Data in Astronomy*

Astronomy has become data intensive. Modern observatories may now generate petabyte-scale data that need new strategies for data management and analysis (**?**). The fourth paradigm enables discoveries from interpreting massive data sets. However, these advances also expose alarming issues, including bottlenecks in the data pipeline, storage challenges, increased skills needed to handle the data, and open access concerns. The field of astronomy is both a beneficiary and a victim of this data-intensive transition.

As mentioned above, the exponential growth of data acquisition can be attributed to Moore's law (**?**). Moore's law predicts that integrated circuit chip density doubles approximately each year at a fixed price point (**?**). (**?**) questioned whether technical development would sustain the growth.

Moore's law can be seen in many data-intensive fields, including astronomy. It explains both the recent development of big data in astronomy, and predicts future challenges.

This paper therefore seeks to review the rise of big data in astronomy and the technical and scientific issues surrounding it by examining four case studies: MeerKAT [1], The Sloan Digital Sky Survey (SDSS) [2], The Legacy Survey of Space and Time (LSST) [3], and The Square Kilometre Array (SKA) [4]. These facilities represent the scope of contemporary astronomical data, the methods of its acquisition, their relative successes, the ongoing challenges, and the solutions currently in use.

[1] https://www.skao.int/en
[2] https://www.sdss.org/
[3] https://www.lsst.org/
[4] https://www.skao.int/en

### 1.3. *An Overview of The Four Surveys*

The SDSS is vital to this paper, as it is one of the earliest large-scale optical surveys that marks the start of the fourth paradigm. The SDSS is a precursor to LSST. The SDSS consists of three main telescopes.

The first of the three is The Sloan Foundation 2.5m Telescope. The Telescope is stationed at the Apache Point Observatory in New Mexico, where it observes the sky in the northern hemisphere. It is able to observe a 3° field of view by use of two corrector lenses (**?**).

The SDSS also uses the Irénée du Pont telescope at Las Campanas Observatory [5]. This telescope is stationed in Chile, where it observes the southern hemisphere instead. Similar to the foundational telescope at Apache Point, this telescope has a 2.1° field of view but only uses one corrector lens (**?**).

The third telescope is the NMSU 1-meter Telescope [6]. The NMSU telescope is stationed at the Apache Point Observatory alongside the foundational telescope. The NMSU telescope is designed to observe bright stars that are too bright for the aforementioned two telescopes to observe (**?**).

the SDSS is made up of multiple subsurveys. The eBoss survey [7], a continuation of BOSS, uses spectrographs to observe light in a wavelength range of 3600-10,400 Å (**?**). An additional subsurvey is APOGEE-2, a continuation of APOGEE. It uses spectrographs similar to eBOSS, but APOGEE-2 collected near-infrared spectra (**?**). MaNGA [8] is a subsurvey that collects integral field unit spectra of 10,000 nearby galaxies (**?**). MARVELS [9] is another SDSS subsurvey, it was built specifically to obtain radial velocity measurements of stars with high-precision in hopes of finding exoplanets (**?**).

The MeerKAT [10] is an important precursor telescope to the SKA (**?**) MeerKAT became fully operational in 2018 in the Northern Cape Province of South Africa. MeerKAT comprises 64 antennas distributed over a radius of approximately 600 miles (**?**). These antennas operate across frequency bands ranging from 350 MHz to 3500 MHz (**?**).

MeerKAT has conducted and continues to conduct ten major survey projects (**?**). For conciseness, this discussion will focus on five of these surveys. One is the LADUMA [11] survey. The objective of the LADUMA survey is to use HI obversations to research galaxy evolution over approximately 9.8 billion years (**?**). LADUMA has used MeerKAT's Phase 1 receivers, which cover 0.9-1.75 GHz. It later transitioned to longer observations in Phase 4, which cover the 0.58-2.5 GHz band (**?**). Although the LADUMA survey is still ongoing, a portion of the data has already been released and will be discussed in the Methods section.

The MeerKAT absorbtion line survey [12] (MALS) is a survey of HI and OH absorbers at a redshift of $z < 0.4$ and $z < 0.7$. HI is a descriptive tracer of the cold neutral medium in a galaxy (**?**). The cold neutral medium contains the physical conditions of the interstellar medium of each galaxy. This, in turn, allows scientists to estimate star formation rate in the galaxy (**?**).

[5] https://www.lco.cl/irenee-du-pont-telescope/
[6] https://newapo.apo.nmsu.edu/
[7] https://www.sdss4.org/surveys/eboss/
[8] https://www.sdss4.org/surveys/manga/
[9] https://www.sdss4.org/surveys/marvels/
[10] https://www.sarao.ac.za/science/meerkat/
[11] https://science.uct.ac.za/laduma
[12] https://mals.iucaa.in/

Another survey, ThunderKAT [13], aims to find, identify and understand high-energy radio transients, usually grouped with observations at similar wavelengths. Examples include supernovae, microquasars, and similar events (**?**).

Another notable MeerKAT survey is MHONGOOSE [14]. This survey aims to catalogue the properties of HI gas using 30 nearby star-forming spiral and dwarf galaxies. MHONGOOSE is remarkable for its higher sensitivity compared to previous surveys such as HALOGAS [15] and THINGS [16] (**?**). This sensitivity is crucial for investigating how low-column-density gas influences the cosmic web and galactic accretion processes (**?**).

The final MeerKAT survey considered here is MIGHTEE [17]. MIGHTEE spans 900-1670 MHz, achieving a resolution of approximately 6 arcseconds. MIGHTEE seeks to study the evolution of active galatic nuclei, neutral hydrogen, and the properties of cosmic magnetic fields (**?**).

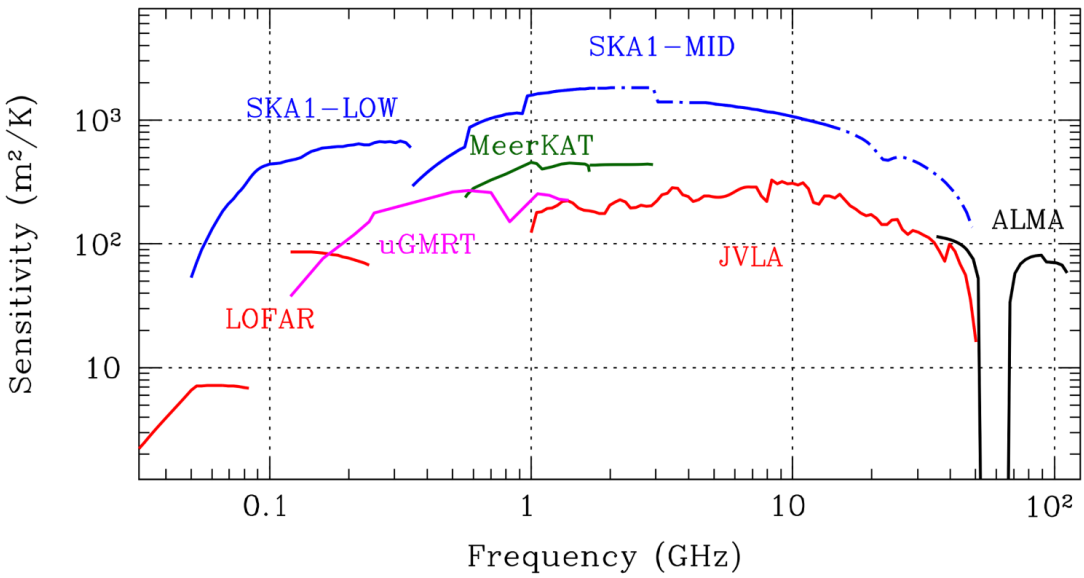


**Figure 1.** Figure from the SKA Official website, demonstrating the sensitivity compared to similar observatories

The SKA has built on technical and scientific achievements paved by MeerKAT and other radio interferometers. The SKA covers an area of approximately 131,205 antennas (**?**). The SKA represents the start of a new frontier for big data in astronomy. As an interferometer it uses aperture synthesis, which allows for the signals from antennas to be phased, this allows to reduce noise (**?**). The SKA will be discussed in further detail in the Methods section.

Alongside the SKA is its optical big data counterpart, the LSST. As noted above, the LSST is a successor to the SDSS. Rubin/LSST, however, has much more sophisticated goals. The LSST plans to address four key scientific issues:Investigating dark energy and dark matter, cataloguing the solar system, collecting data for sky surveys, and mapping the Milky Way. To achieve this, the LSST uses a 3.2-gigapixel camera with a sampling of 9.6 deg$^2$ field of view (**?**). These cameras are equipped

[13] https://www.physics.ox.ac.uk/research/group/meerkat
[14] https://mhongoose.astron.nl/
[15] https://www.astron.nl/halogas/
[16] https://www2.mpia-hd.mpg.de/THINGS/Overview.html
[17] https://www.mighteesurvey.org/home

with highly resistant sensors reinforced with silicon (**?**). Rubin/LSST has an unprecented data rate for an optical telescope.

The SDSS, MeerKAT, SKA, and LSST generate unprecedented data rates and allow the experimentation of complex astrophysical events and phenomena. In the following Methods section, I describe how the data are collected, processed, analyzed, and stored. I then compare SDSS and MeerKAT to their larger successor telescopes, Rubin/LSST and SKA and consider the evolution of data challenges.

## 2. METHODS

### 2.1. *The Optical Big Data Pipeline*

This section carefully examines how each of the optical focused surveys collects and processes its data. It begins by describing the nature, scope, and type of the data. Next, it discusses how each survey collecs and archives its data, followed by an explanation of their general data processing methods. Finally, it considers the use of real-time data processing. By comparing these surveys, this study highlights the rapid growth of big data in astronomy, a trend that has created challenges for data storage, processing, and analysis. These challenges will be discussed further in the discussion section.

The first survey examined is the Sloan Digital Sky Survey (SDSS). The SDSS I plan on splitting the SDSS into its three main components, the 2.5 m Telescope, the Irénée du Pont Telescope, and the NMSU 1-meter telescope. As discussed previously, I will explain how each telescope collects data, then I will explain how the SDSS processes the data both generally and in real-time. Lastly, I will talk about the open data policy the SDSS has employed.

#### 2.1.1. *The Sloan Digital Sky Survey (SDSS)*

*Data Collection and Storage*—Although the SDSS is a complex survey, it can be divided into several major components, each of which contributes to the collection of astrophysical data. The 2.5-meter telescope plays a central role in the operations of the SDSS. It was designed to conduct precise optical observations of the sky over many years.

**(1) The SDSS 2.5 m Telescope:** According to Gunn et al. (2006), the SDSS camera contains "30 2048 x 2048 Scientific Imaging Technologies Charge-Coupled Devices (CCDs) and 24 2048 x 400 CCDs" (**?**). A CCD is a detector that converts incoming light into an electronic signal. When photons strike the CCD, they generate electrons through the photoelectric effect. Using applied voltages, the resulting charge is measured based on the number of electrons produced. That measurement is then converted into a digital value and stored as a pixel, forming an image (**?**).

Another major innovation that enables the SDSS to collect data is the pair of fiber-fed double spectrographs, which record imaging data across wavelengths from 3800 to 9200 Å and at field angles between 0 and 90° (**?**). present measurements of the optical performance of these instruments, which are summarized in Figure 2.

In Figure 5, $\lambda$ represents the wavelength of light, and "Angle" refers to the field angle. $f_b$ is the best-focus distance, which is the position that provides the sharpest image. $h/dh$ represents the lateral color, and $D$ indicates the longitudinal difference from the best focus. Finally, $\epsilon$ is the root mean square (rms) image diameter. Among these parameters, the lateral color and longitudinal

| $\lambda$ (Å) | Angle (arcmin) | $f_h$ (mm) | $h/dh$ (mm) | $D$ (mm) | $\epsilon$ (mm) |
|---|---|---|---|---|---|
| 4000 | 0.00 | −0.007 | 0.000 | −0.135 | 0.036 |
| | 30.00 | −0.143 | 0.004 | −0.081 | 0.030 |
| | 45.00 | −0.424 | 0.005 | −0.015 | 0.025 |
| | 60.00 | −0.978 | 0.005 | 0.076 | 0.028 |
| | 70.00 | −1.536 | 0.004 | 0.148 | 0.036 |
| | 80.00 | −2.265 | 0.002 | 0.231 | 0.049 |
| | 90.00 | −3.203 | −0.004 | 0.325 | 0.065 |
| 4600 | 0.00 | −0.007 | −0.000 | −0.058 | 0.031 |
| | 30.00 | −0.143 | 0.002 | −0.035 | 0.027 |
| | 45.00 | −0.424 | 0.002 | −0.006 | 0.024 |
| | 60.00 | −0.978 | 0.002 | 0.033 | 0.025 |
| | 70.00 | −1.536 | 0.002 | 0.065 | 0.027 |
| | 80.00 | −2.265 | 0.001 | 0.101 | 0.030 |
| | 90.00 | −3.203 | −0.001 | 0.141 | 0.035 |
| 5300 | 0.00 | −0.007 | 0.000 | 0.000 | 0.029 |
| | 30.00 | −0.143 | −108.818 | 0.000 | 0.026 |
| | 45.00 | −0.424 | −163.322 | 0.000 | 0.024 |
| | 60.00 | −0.978 | −217.855 | 0.000 | 0.025 |
| | 70.00 | −1.536 | −254.241 | 0.000 | 0.027 |
| | 80.00 | −2.265 | −290.713 | 0.000 | 0.026 |
| | 90.00 | −3.203 | −327.372 | 0.000 | 0.025 |
| 6500 | 0.00 | −0.007 | 0.000 | 0.062 | 0.031 |
| | 30.00 | −0.143 | −0.002 | 0.037 | 0.027 |
| | 45.00 | −0.424 | −0.002 | 0.007 | 0.024 |
| | 60.00 | −0.978 | −0.002 | −0.035 | 0.029 |
| | 70.00 | −1.536 | −0.002 | −0.068 | 0.034 |
| | 80.00 | −2.265 | −0.001 | −0.106 | 0.036 |
| | 90.00 | −3.203 | 0.002 | −0.149 | 0.040 |
| 9000 | 0.00 | −0.007 | 0.000 | 0.131 | 0.036 |
| | 30.00 | −0.143 | −0.004 | 0.078 | 0.029 |
| | 45.00 | −0.424 | −0.004 | 0.014 | 0.026 |
| | 60.00 | −0.978 | −0.004 | −0.074 | 0.036 |
| | 70.00 | −1.536 | −0.004 | −0.145 | 0.046 |
| | 80.00 | −2.265 | −0.002 | −0.226 | 0.056 |
| | 90.00 | −3.203 | 0.003 | −0.317 | 0.068 |

**Figure 2.** Figure 5 from Gunn et al. (2006), showing results of the SDSS spectrographs given Wavelength and Angle.

difference are the most important for image quality because smaller values indicate sharper images. Based on the data from (**?**), both of these quantities remain close to zero for most wavelengths and field angles, except between roughly 5300 and 6500 Å, which demonstrates the high optical accuracy of the SDSS spectrographs.

The combination of these two innovations alongside others help the 2.5 m telescope collect data at a rate of about 20Gb/hr (**?**).

**(2) The Irénée du Pont Telescope:** Unlike the 2.5 m telescope, the Du Pont Telescope does not rely on CCDs to collect data. According to Bowen's 1973 paper, the telescope is described as a modified Ritchey-Chrétien design with Gascoigne correctors (**?**). The Du pont telescope uses a 100-inch primary mirror. Approximately 40% of the light is reflected to the secondary mirror, obtaining only a 16% loss of light at that stage. (**?**) The combination of light from the two aforementioned mirrors are then sent to a 20 inch x 20 inch plate, where monocromatic images are formed.

The du Pont Telescope uses 18.9 inch nonvignetted plates in order to minimize vignetting (**?**). Vignetting is the process where light beds through the lense of a telescope. The bending form a cone of light, which causes images to be darker near the edges and brighter in the center of the image (**?**). Because of the nonvignetted plates, the du Pont Telescope experiences an exceptionally low 3% percent loss of light (**?**).

Another technology the du Pont Telescope applies is a Gascoigne corrector plate. The plate helps with data collection. The Gasciogne corrector plate is able to be moved, which can help optimize the

collection of light in a wanted wavelength (**?**). Given a seperation of 1000 mm from the end of the corrector plate to the focus gives an image with a minimized astigmatism for a refractive index of n = 1.47 (**?**). At a given wavelength, the change of length which minimizes astigmatism is described in Bowen's paper as

$$\Delta L = 590\Delta n/(n-1) = -1250\Delta n \tag{1}$$

Where $\Delta L$ is the change in seperation in millimeters and $\Delta n$ is the difference between a refractive index of 1.47 and the index wanted.

The last technology the du Pont Telescope uses is conical baffles. The reason for this is to promote shielding in the telescope (**?**). As explained in the Bowen paper, shielding is necessary in order to protect the photographic plate from light that escapes from the secondary lense due to long time exposure. As explained in the Bowen paper, the conic baffles are "located in the space between the incoming beam as it appraoches the primary and the return beam from the secondary to the plate" (**?**). Theoretically, the conic baffles have the disatvantage of producing a diffraction pattern. However, as explained by Bowen, this should not majorly affect the images of stars (**?**).

**(3) The New Mexico State University (NMSU) Telescope:** The NMSU telescope takes the most technologically advanced approach to collecting data compared to the 2.5 m telescope and the du Pont Telescope. The NMSU telescope uses a camera that has a 2048 x 2048 CCD. The camera is controled by a linux computer, which is connected by fiber optic cables (**?**).

The data collection of the NMSU telescope is almost fully automated using C++, the only time it is not is when engineering is being done on the telescope (**?**). The NMSU telescope has a camera which analyzes the brightness level of the sky to see if it is dark enough to start collecting data. When the sky becomes dark enough, the telescope initiates its program. The telescope goes through the list and observes said objects (**?**). Objects can, however, be observed as many times as requested.

*Data Processing*—The SDSS processes its data through an innovative acquisition system that records and organizes observations in real time while maintaining strict quality control (**?**). This system ensures that data are processed and stored efficiently without any loss of precision. (**?**). The data pipeline of the SDSS can be described by two different fields of data, the imaging pipeline and the spectroscopy pipeline

**(1) Imaging Data Pipeline:** The Imaging data Pipeline itself consists of multiple subpipelines. The first subpipeline is the Astroline. The astroline uses vxWorks in order to initalize the processing sequence. This happens by composing star cutouts and column quartiles collected from the CCD's mentioned before (**?**)

The second subpipline is the MT pipeline. the MT Pipeline processes the data collected from the Photometric Telescope. This data is used to calculate important parameters for the 2.5 m Telescope scans, such as extinction and zero-points (**?**).

The third pipeline is the Serial Stamp Collecing (SSC) Pipeline. the SSC reorganizes the star cutouts collected from previous pipelines. The SSC does this in order to prepare data for the upcoming pipelines (**?**).

Next is the Astrometric Pipeline. The Astrometric pipeline processes the average location of stars using the data collected from the astroline and SSC pipelines. It then converts the pixel data from the images into celestial coordinates $(\alpha, \delta)$ (**?**).

After that is the Postage Stamp Pipeline (PSP). The PSP estimates the quality of the data collected by calculating factors such as the flat field vectors, bias drift, and sky levels (**?**).

After all that is done, the data is fed into the Frames Pipeline. The Frames pipeline does a majority of the work, processing the data from all the previous pipelines and producing the complete datasets of images. It does this by correcting the frames based on the data before and cataloging the images (**?**).

Lastly, the processed data is then ran through the Calibration pipeline. The Calibration pipeline takes data from the MT and Frames pipeline. The Calibration pipeline converts the counts into more measurable quantities such as flux (**?**).

The SDSS imaging pipeline is composed of multiple connected pipelines that operate collaboratively to transform raw imaging data into structured datasets from which measurable physical quantities such as flux can be derived.

**(2) Spectroscopy Data Pipeline:**

*Real-Time Processing—*

### 2.1.2. *The Rubin/Large Synoptic Survey Telescope (LSST)*

*Data Collection and Storage*—Despite the recent times in creation. The SDSS collected around 16 TB of data over a decade in their data release 7 (**?**). Yet the LSST is expected to collect 20 TB of data per night (**?**).

The LSST pipeline is written with around 750000 in Python in order to use relevant libraries such as SciPy and AstroPy (**?**). The pipeline is also written with around 220000 lines in C++ in order to ensure efficient performance (**?**). In order to combine these two languages, pybind11 allows for the transition from Python to C++, and ndarray objects are able to be converted from C++ arrays (**?**).

The python enviroment of the LSST pipeline uses a package named `rubin-env`. This package gives the user all the code needed to run LSST's data. In order to execute the code, the pipeline consists multiple packages that each serve their own purpose. The LSST has defined a class labled `Task`, which is used to define algorithms (**?**).

One instance of a task is the `PipelineTask`, which serves to organize subtasks. These subtasks each have their own purpose (**?**). The most important subtask, labeled `daf_butler`, handles the data storage. This subtask is titled by the LSST as The Data Butler. The Butler serves as a database to store collected data. It stores objects with data IDs similar to SQL, with headers that hold useful information (**?**). An example of this would be a data coordinate labeled `instrument`="LSSTCam", `exposure`=299792458, `detector`=42, `band`=z, `day_obs`=20251011.

*Data Processing*—There are multiple tasks which define how the LSST processed data to find objects. These are all defined in the `meas_algorithms` package (**?**). The task that first handles processing the catalogued images is the `SourceDetectionTask` (**?**). This task uses Gaussian smoothing in the point spread function. It then convolves the collected image with the point spread function in order to suppress potential noise (**?**).

Another task that the LSST uses to process data is `MaskStreaksTask` (**?**). The task serves to mask pixels from streaks from other satellites. It identifies streaks using a Canny Filter and the Kernel-Based Hough Transform (**??**). This task is combined with the deblending of collected images allows for the LSST to accurately identify objects.

*Real-Time Processing*—

## 2.2. *The Radio Big Data Pipeline*

### 2.2.1. *The MeerKAT*

**NEW TEXT STARTS HERE:**

*Data Processing*—The MeerKAT data processing pipeline is split into three parts, the calibration pipeline, the continuum pipeline, and the spectral pipeline (**?**).

**(1) The Calibration Pipeline:** The pipeline starts by choosing an antenna as a reference point for the other antennas (**?**). This process is done by collecting the first scan of an antenna. The data is then averaged and fourier transformed. After that, it is the ratio called the peak-to-noise ratio is measured. The peak-to-noise ratio is the ratio of the maximum of the fourier transformed data and the rms noise at a distance far from the peak (**?**). After tha the process repeats for every antenna and the one with the highest median with respect to every baseline for the peak-to-noise ratio is chosen as the reference antenna (**?**). This antenna will have all of its calibration solutions set to zero (**?**).

The purpose of the calibration pipeline is to ensure that the reference antenna and the other antennas by extension are callibrated. At the start of every observation, the reference antenna is evaluated based on certain flags, such as data loss (**?**). If 80% or more flags are pinged, a new reference antenna is determined based on the aforementioned process (**?**).

**(2) The Continuum Pipeline:** The second pipeline is the continuum pipeline. This pipeline aims to produce continuum images using the orbit software package (**??**). The orbit package which interfaces UV data. It does this by using the `MFImage` task, which does wide-band, wide-field imaging (**??**). Once the data is read, it is split into scans, which are then averaged and combined into a dataset.

This dataset is then split into 8 107 MHz intermediate frequencies (**?**). The UV data then stores headers of `ntimes, nbaselines, nchannels, npolarisation` (**?**). Given the Meerkat's design of 48 antennas, short baselines dominate. This causes an inflation in data volume (**?**). In order to circumvent this, a baseline dependent averaging step is applied (**?**). Time averaging the data with respect to the baseline before merging them into a dataset. This step reduces the data volume by about 3-4 times with a minimal loss of 1

The aforementioned `MFImage` task uses joint frequency deconvolution to handle wide-band effects in wide-band images (**?**). Normally, Meerkat uses around 140 circular facets with a size of around 6 arcminutes which cover 1 degree from the phase center (**?**). Additionally, there are facets with around 1 degree that use the SUMSS or NVSS catalogue for a radius of 2.5 from the center of the phase (**?**). These 1 degree facets are used for phenomena that are ancipiated to have a flux density greater than 5 mJy (**?**). Due to the fact that these facets deal with wide-band images, they also deal with wide-band imaging effects. These effects are bypassed by splitting the frequency band used into 10 components (**?**). When the joint frequency deconvolution is initiated, the brightest light sources in

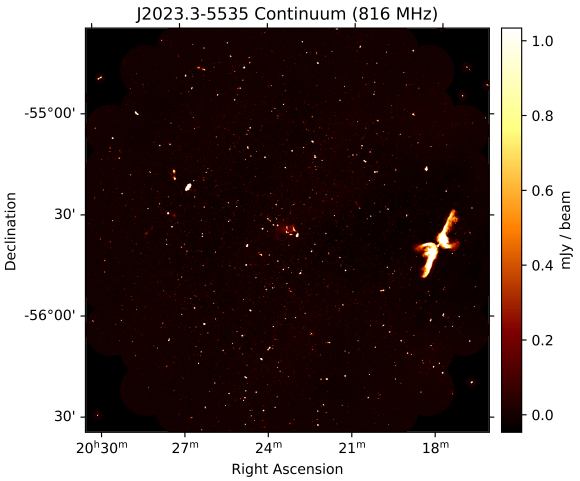


**Figure 3.** An example UHF continuum thumbnail image produced automatically by the pipeline is shown above, (**?**)

the dataset are found and subtracted from the slices of data individually. This is called the CLEAN algorithm (**?**).

Self-calibration is then performed in two rounds in order to transform data. The first round consists of using the CLEAN algorithm with a spectral density of 1 mJy. This yields approximately 1000 CLEAN components (**?**). The data is then self calibrated once more, using a CLEAN algorithm with a spectral density of 100 $\mu$Jy. The second round yields approximately 10000 CLEAN components (**?**). After this process, the data is imaged and used to create the continuum images. Lastly, the self calibrated data is converted into AIPS format, which consists of tables with headers such as UV data, polarisation, etc (**?**). The data is then converted into numpy arrays for each timestamp with shape (nif, npol, nantennas) (**?**). The sky model, made up of the CLEAN components, is stored in AIPS CC format. The flux density of all the 10 components of the frequency are summated (**?**). The merged flux density is then fitted with a second degree polynomial to measure frequency and is used to subtract from the image (**?**). Finally, the fully processed images are converted into FITS and PNG files and archived (**?**).

**(3) The Spectral Pipeline:** The purpose of the spectral pipeline is to speed up image creation while keeping control of quality (**?**). The issue with the imaging data, at this point is that they're independent of eachother, which causes issues with accessing data when trying to run said images (**?**). The raw data, also known as visibilities. are recieved in time-major order, but this data structure must be transposed in order for the data to be in channel-major order (**?**). In order to solve this issue, a visibility writer is used. The writer stores the visibilities on a Ceph cluster with chunks each across 64 channels (**?**). This choice in the chunk size is not optimized, but it does avoid issues with RAM and memory allocation (**?**).

The chunk system only partially implements the solution of transposing the data. The solution is complete after another series of events are done (**?**). The visibilities in each chunk are ordered by channel, W slice, and baseline. This allows for both the accessibility of the data to improve, and it inherently stores measurements such as UVW coordinateas and parallatic angles (**?**). While all that is occuring, conservative baseline-depedent averaging is also applied to the visibilities (**?**). The

coordinates of every visibility are solved for, then visibilities with matching coordinates are merged, which is the last of the preprocessing of the visiblities (**?**).

After all of that is done, By using CUDA, a parallel processing software, every channel is imaged separetly. The use of CUDA allows for the processing to speed up due to the usage of NVIDIA GPUs (**?**). Lastly, the data follows a cycle in order to be updated as described in the figure below.

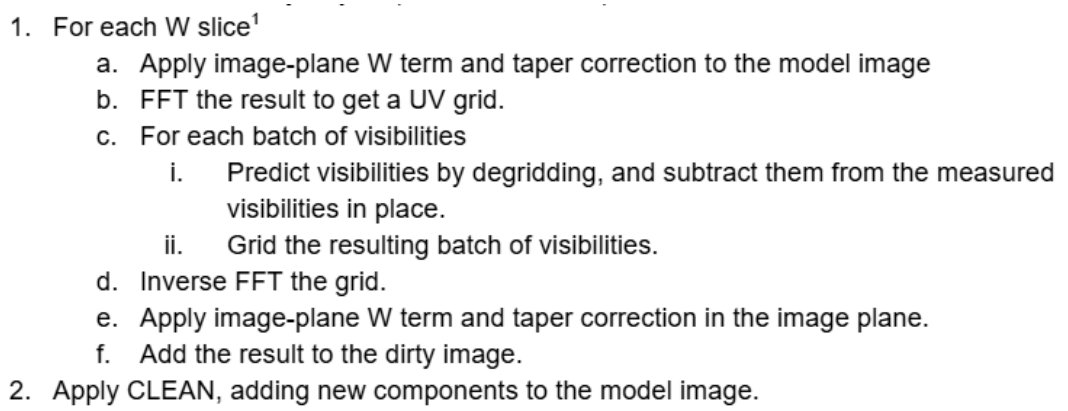


**Figure 4.** The cyclic algorithm for the Meerkat's spectrographic pipeline, (**?**)

NEW TEXT ENDS HERE:

#### 2.2.2. *The Square Kilometre Array (SKA)*

## 3. RESULTS

### 3.1.

## 4. DISCUSSION

### 4.1. *Open Source Policies and Transparency*

#### 4.1.1. *The SDSS Policy*

The SDSS collaboration states on its official SDSS-IV website [18] that all of its software be open source using the open source liscence BSD 3-Clause. However, the SDSS outlines practices for users who plan to use the SDSS software must abide by. One of the most important ones is the proper citation of software and websites that were used. The SDSS4 emphasizes the importance of citing properly as it serves to acknowledge the hard work of the teams behind said projects.

The SDSS also has implemented Digital Object Identifiers, commonly known as DOIs, into all software code. DOIs allow for software and data to be easily identified, which is important for ownership. The SDSS team also promotes transparency in coding by implementing Git and SVN in order to maintain a record of the development of the software. This not only makes the development transparent, but also helps users see the evolution of the software.

[18] https://www.sdss4.org/dr17/software/

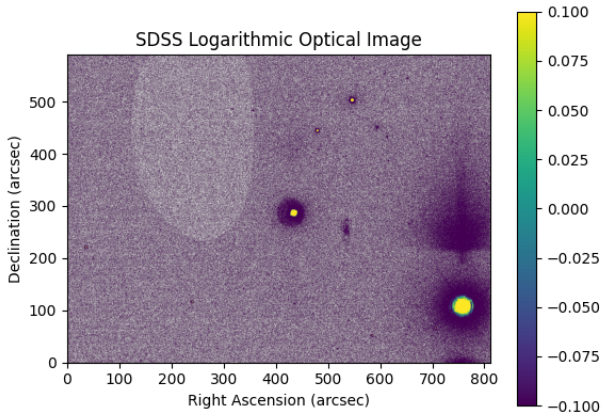


**Figure 5.** A spectrographical image obtained using data collected from SDSS

Overall, the SDSS has demonstrated a strong commitment to making their data and software open source and transparent. This in turn helps the development of science, by ensuring that knowledge is accessible to all regardless of resources.

#### 4.1.2. *The LSST Policy*

**New text starts here:**

#### 4.1.3. *The MeerKAT Policy*

**New text ends here:**

#### 4.1.4. *The SKA Policy*

## A. PYTHON CODE FOR SDSS DATA RETRIEVAL (FIGURE 3)

```
#Import relevant libraries/functions
from astroquery.sdss import SDSS
from astropy import coordinates as coords
import astropy.units as u
import matplotlib.pyplot as plt
import numpy as np

#Initalize Right Ascension and Declination
ra = 20
dec = -10

#Convert ra and dec into a SkyCoord Object
coord = coords.SkyCoord(ra, dec, unit='deg', frame = 'icrs')

#Query the SDSS System to find object given coordinates in a radius of
    0.01 degrees
result = SDSS.query_region(coord,  radius=0.01*u.deg, spectro=True)
```

```
print(result)
#Retrieve the Image from found object, make into a FITS File
image = SDSS.get_images(matches=result, band=['u', 'g', 'r', 'i', 'z'])


#Retrieve hdulist from FITS file
hdulist = image[0]

#Retrieve the image data from the hdulist
imageData = hdulist[0].data

#Log the image data in order to get rid of background
imageDataLog = np.log10(imageData) + 1e-8

#Save the header
header = hdulist[0].header


#Obtain the relevant headers

#Retrieve pixel scale numbers, divided amongst two parts for ra and dec
CD1_1 = header['CD1_1']
CD1_2 = header['CD1_2']
CD2_1 = header['CD2_1']
CD2_2 = header['CD2_2']

#Width of image in pixels
keyWordNAXIS1 = header['NAXIS1'] #[pixels]

#Height of image in pixels [pixels]
keyWordNAXIS2 = header['NAXIS2'] #[pixels]

#Normalize the pixel scale, then multiply by 3600 to convert units
CDELT1ArcSec = np.linalg.norm([CD1_1,CD1_2]) * 3600 #[arcsec/pixels]
CDELT2ArcSec = np.linalg.norm([CD2_1,CD2_2]) * 3600 #[arcsec/pixels]

#Set up the image
plt.xlabel("Right Ascension (arcsec)")
plt.ylabel("Declination (arcsec)")
plt.title('SDSS Logarithmic Optical Image')
vmin2 = np.percentile(imageDataLog, 85)
vmax2 = np.percentile(imageDataLog, 98)
plt.imshow(imageDataLog, cmap='viridis', extent = (0, CDELT1ArcSec *
   keyWordNAXIS1, 0, CDELT2ArcSec * keyWordNAXIS2), vmin = vmin2, vmax =
   vmax2)
plt.colorbar()

plt.show()
```

## REFERENCES


????, The MIGHTEE Survey, https://www.mighteesurvey.org/home

????, SKA Telescope Specifications, https://www.skao.int/en/science-users/118/ska-telescope-specifications

Blyth, S., Baker, A. J., Holwerda, B., et al. 2018, in Proceedings of MeerKAT Science: On the Pathway to the SKA — PoS(MeerKAT2016) (Stellenbosch, South Africa: Sissa Medialab), 004, doi: 10.22323/1.277.0004

Bowen, I. S., & Vaughan, A. H. 1973, Applied Optics, 12, 1430, doi: 10.1364/AO.12.001430

Bundy, K., Bershady, M. A., Law, D. R., et al. 2014a, The Astrophysical Journal, 798, 7, doi: 10.1088/0004-637X/798/1/7

—. 2014b, The Astrophysical Journal, 798, 7, doi: 10.1088/0004-637X/798/1/7

Camilo, F. 2024

Cotton, W. D. 2008, Publications of the Astronomical Society of the Pacific, 120, 439, doi: 10.1086/586754

Cotton, W. D., & Schwab, F. R. 2010

Dawson, K. S., Kneib, J.-P., Percival, W. J., et al. 2016, The Astronomical Journal, 151, 44, doi: 10.3847/0004-6256/151/2/44

De Blok, W. J. G., Healy, J., Maccagni, F. M., et al. 2024, Astronomy & Astrophysics, 688, A109, doi: 10.1051/0004-6361/202348297

Dewdney, P., Hall, P., Schilizzi, R., & Lazio, T. 2009, Proceedings of the IEEE, 97, 1482, doi: 10.1109/JPROC.2009.2021005

Fernandes, L. A., & Oliveira, M. M. 2008, Pattern Recognition, 41, 299, doi: 10.1016/j.patcog.2007.04.003

Goedhart, S. 2025, MeerKAT Specifications

Gunn, J. E., Siegmund, W. A., Mannery, E. J., et al. 2006, The Astronomical Journal, 131, 2332, doi: 10.1086/500975

Gupta, N., Jagannathan, P., Srianand, R., et al. 2021, The Astrophysical Journal, 907, 11, doi: 10.3847/1538-4357/abcb85

Hey, T., Tansley, S., & Tolle, K. 2009, Microsoft Research

Holtzman, J. A., Harrison, T. E., & Coughlin, J. L. 2010, Advances in Astronomy, 2010, 193086, doi: 10.1155/2010/193086

Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, The Astrophysical Journal, 873, 111, doi: 10.3847/1538-4357/ab042c

Jonas, J., & the MeerKAT Team. 2018, in Proceedings of MeerKAT Science: On the Pathway to the SKA — PoS(MeerKAT2016) (Stellenbosch, South Africa: Sissa Medialab), 001, doi: 10.22323/1.277.0001

Juric, M., Kantor, J., Lim, K.-T., et al. 2017

Lesser, M. 2015, Publications of the Astronomical Society of the Pacific, 127, 1097, doi: 10.1086/684054

Lupton, R., Gunn, J. E., Ivezic, Z., et al. 2001, The SDSS Imaging Pipelines, arXiv, doi: 10.48550/arXiv.astro-ph/0101420

Lupton, R. H., Ivezic, ˇ., Gunn, J., et al. 2007

Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., et al. 2017, The Astronomical Journal, 154, 94, doi: 10.3847/1538-3881/aa784d

Moore, G. E. 2006, IEEE Solid-State Circuits Society Newsletter, 11, 33, doi: 10.1109/N-SSC.2006.4785860

NSF-DOE Vera C. Rubin Observatory. 2025, PSTN-019: The LSST Science Pipelines Software: Optical Survey Pipeline Reduction and Analysis Environment, NSF-DOE Vera C. Rubin Observatory, doi: 10.71929/RUBIN/2570545

Ratcliffe, S. 2021, SDP Pipelines Overview, https://skaafrica.atlassian.net/wiki/spaces/ESDKB/pages/

Richards, S. 2020, What Is Vignetting?

Verbunt, F., & Van Gent, R. H. 2010, Astronomy and Astrophysics, 516, A28, doi: 10.1051/0004-6361/201014002

Woudt, P. A., Fender, R., Corbel, S., et al. 2018, in Proceedings of MeerKAT Science: On the Pathway to the SKA — PoS(MeerKAT2016) (Stellenbosch, South Africa: Sissa Medialab), 013, doi: 10.22323/1.277.0013

# REFERENCES


????, The MIGHTEE Survey, https://www.mighteesurvey.org/home

????, SKA Telescope Specifications, https://www.skao.int/en/science-users/118/ska-telescope-specifications

Blyth, S., Baker, A. J., Holwerda, B., et al. 2018, in Proceedings of MeerKAT Science: On the Pathway to the SKA — PoS(MeerKAT2016) (Stellenbosch, South Africa: Sissa Medialab), 004, doi: 10.22323/1.277.0004

Bowen, I. S., & Vaughan, A. H. 1973, Applied Optics, 12, 1430, doi: 10.1364/AO.12.001430

Bundy, K., Bershady, M. A., Law, D. R., et al. 2014a, The Astrophysical Journal, 798, 7, doi: 10.1088/0004-637X/798/1/7

—. 2014b, The Astrophysical Journal, 798, 7, doi: 10.1088/0004-637X/798/1/7

Camilo, F. 2024

Cotton, W. D. 2008, Publications of the Astronomical Society of the Pacific, 120, 439, doi: 10.1086/586754

Cotton, W. D., & Schwab, F. R. 2010

Dawson, K. S., Kneib, J.-P., Percival, W. J., et al. 2016, The Astronomical Journal, 151, 44, doi: 10.3847/0004-6256/151/2/44

De Blok, W. J. G., Healy, J., Maccagni, F. M., et al. 2024, Astronomy & Astrophysics, 688, A109, doi: 10.1051/0004-6361/202348297

Dewdney, P., Hall, P., Schilizzi, R., & Lazio, T. 2009, Proceedings of the IEEE, 97, 1482, doi: 10.1109/JPROC.2009.2021005

Fernandes, L. A., & Oliveira, M. M. 2008, Pattern Recognition, 41, 299, doi: 10.1016/j.patcog.2007.04.003

Goedhart, S. 2025, MeerKAT Specifications

Gunn, J. E., Siegmund, W. A., Mannery, E. J., et al. 2006, The Astronomical Journal, 131, 2332, doi: 10.1086/500975

Gupta, N., Jagannathan, P., Srianand, R., et al. 2021, The Astrophysical Journal, 907, 11, doi: 10.3847/1538-4357/abcb85

Hey, T., Tansley, S., & Tolle, K. 2009, Microsoft Research

Holtzman, J. A., Harrison, T. E., & Coughlin, J. L. 2010, Advances in Astronomy, 2010, 193086, doi: 10.1155/2010/193086

Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, The Astrophysical Journal, 873, 111, doi: 10.3847/1538-4357/ab042c

Jonas, J., & the MeerKAT Team. 2018, in Proceedings of MeerKAT Science: On the Pathway to the SKA — PoS(MeerKAT2016) (Stellenbosch, South Africa: Sissa Medialab), 001, doi: 10.22323/1.277.0001

Juric, M., Kantor, J., Lim, K.-T., et al. 2017

Lesser, M. 2015, Publications of the Astronomical Society of the Pacific, 127, 1097, doi: 10.1086/684054

Lupton, R., Gunn, J. E., Ivezic, Z., et al. 2001, The SDSS Imaging Pipelines, arXiv, doi: 10.48550/arXiv.astro-ph/0101420

Lupton, R. H., Ivezic, ˇ., Gunn, J., et al. 2007

Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., et al. 2017, The Astronomical Journal, 154, 94, doi: 10.3847/1538-3881/aa784d

Moore, G. E. 2006, IEEE Solid-State Circuits Society Newsletter, 11, 33, doi: 10.1109/N-SSC.2006.4785860

NSF-DOE Vera C. Rubin Observatory. 2025, PSTN-019: The LSST Science Pipelines Software: Optical Survey Pipeline Reduction and Analysis Environment, NSF-DOE Vera C. Rubin Observatory, doi: 10.71929/RUBIN/2570545

Ratcliffe, S. 2021, SDP Pipelines Overview, https://skaafrica.atlassian.net/wiki/spaces/ESDKB/pages/

Richards, S. 2020, What Is Vignetting?

Verbunt, F., & Van Gent, R. H. 2010, Astronomy and Astrophysics, 516, A28, doi: 10.1051/0004-6361/201014002

Woudt, P. A., Fender, R., Corbel, S., et al. 2018, in Proceedings of MeerKAT Science: On the Pathway to the SKA — PoS(MeerKAT2016) (Stellenbosch, South Africa: Sissa Medialab), 013, doi: 10.22323/1.277.0013