Draft version November 9, 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. Do the main part of the discussion. Talk about the exponential increase of data

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 (Hey et al. 2009), 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 (Hey et al. 2009), 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 (Verbunt & Van Gent 2010). 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 (Hey et al. 2009). 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 (Hey et al. 2009). 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 (Hey et al. 2009), 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 (Hey et al. 2009). 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 (Hey et al. 2009). 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 (Hey et al. 2009). 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 (Hey et al. 2009). Moore's law predicts that integrated circuit chip density doubles approximately each year at a fixed price point (Moore 2006). (Moore 2006) 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.

1.3. *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 (Gunn et al. 2006).

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

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 (Bowen & Vaughan 1973).

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 (Majewski et al. 2017).

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 (Dawson et al. 2016). An additional subsurvey is APOGEE-2, a continuation of APOGEE. It uses spectrographs similar to eBOSS, but APOGEE-2 collected near-infrared spectra (Majewski et al. 2017). MaNGA [8] is a subsurvey that collects integral field unit spectra of 10,000 nearby galaxies (Bundy et al. 2014a). 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 (Bundy et al. 2014b).

The MeerKAT [10] is an important precursor telescope to the SKA (Jonas & the MeerKAT Team 2018) 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 (Goedhart 2025). These antennas operate across frequency bands ranging from 350 MHz to 3500 MHz (Goedhart 2025).

MeerKAT has conducted and continues to conduct ten major survey projects (Jonas & the MeerKAT Team 2018). 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 (Blyth et al. 2018). 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 (Blyth et al. 2018). 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 (Gupta et al. 2021). 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 (Gupta et al. 2021).

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 (Woudt et al. 2018).

[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/
[13] https://www.physics.ox.ac.uk/research/group/meerkat

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] (De Blok et al. 2024). This sensitivity is crucial for investigating how low-column-density gas influences the cosmic web and galactic accretion processes (De Blok et al. 2024).

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 (MIG ????).

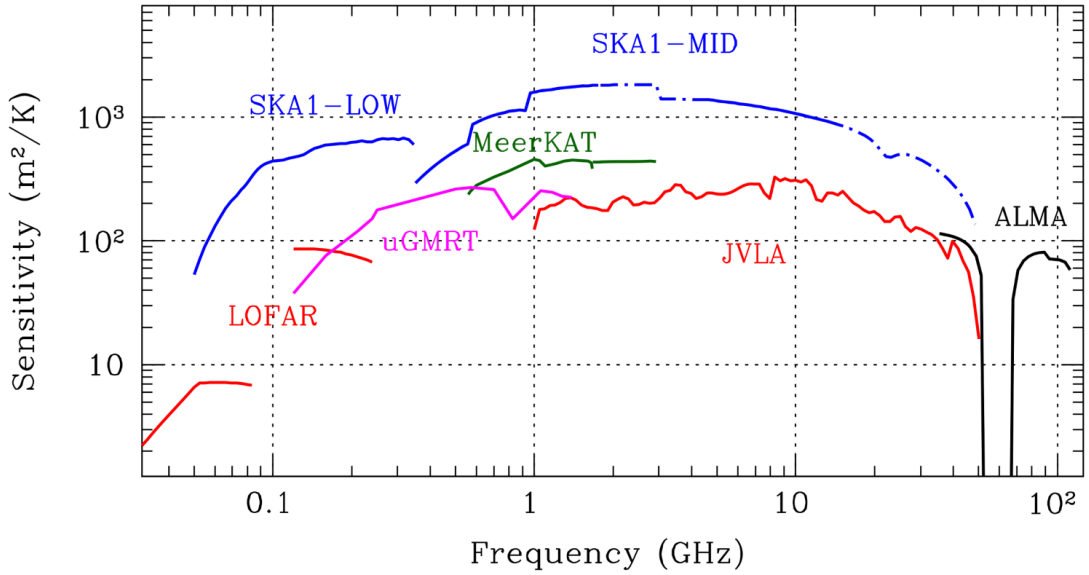

Figure 1. Figure 3 from the SKA Official website [b], SKA1 sensitivity compared to existing facilities at similar frequencies

[a] https://www.skao.int/en/science-users/118/ska-telescope-specifications
[b] https://www.skao.int/en/science-users/118/ska-telescope-specifications

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 (SKA ????). 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 (Dewdney et al. 2009). 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 (Ivezić et al. 2019). These cameras

[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

are equipped with highly resistant sensors reinforced with silicon (Ivezić et al. 2019). 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. By describing the nature, scope, and type of the data. Next, we discuss how each survey collects and archives its data, followed by an explanation of their general data processing methods. Finally, we consider the use of real-time data processing.

As noted Above, we can consider the SDSS to consist of three main components: the 2.5 m foundational telescope, the Irénée du Pont telescope, and the NMSU 1-meter telescope. The SDSS has evolved rapidly overtime. In the early 2000's the SDSS was primarily focused on optical imaging, but ever since the mid 2010's it has shifted towards gathering spectral data (**?**Ahumada et al. 2020) As discussed previously, we will consider how each telescope collects data, how the SDSS processes the data, and makes the data acessible through an open data policy.

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

Data Collection and Storage—**(1) The SDSS 2.5 m Foundational Telescope:** The SDSS camera contains 54 2048 x 2048 charge-coupled devices (CCDs) and 24 2048 x 400 CCDs. A CCD is an imaging detector that converts incoming light into an electronic signal. When photons strike the CCD, they generate electrons through the internal photoelectric effect. The acculumated charge is measured per pixel and is stored as a digital value (Lesser 2015).

In addition to the CCD imaging data the SDSS collects spectra using a pair of fiber-fed double spectrographs, over a wavelength range from 3800 to 9200 Å and at field angles between 0 and 90° (Gunn et al. 2006). The spectrograph fibers are positioned in pre-selected objects of the field. The optical performance of these spectrographs, which are summarized in Table 5 from Gunn et al. (2006).

In Figure 5, λ represents wavelength, and "Angle" refers to the field angle. f_b denotes the best-focus distance. h/dh represents the lateral color, D denotes the longitudinal difference from the best focus, and ϵ is the root mean square (rms) image diameter. Smaller values of the lateral color and longitudinal difference indicate sharper images. Both of these quantities remain close to zero for most wavelengths and field angles, except between roughly 5300 and 6500 Å(Gunn et al. 2006). This demonstrates the high optical accuracy of the SDSS spectrographs. The 2.5 m telescope collects imaging and spectroscopic data at a rate of about 20Gb/hr (Lupton et al. 2007).

(2) The Irénée du Pont Telescope: The du Pont telescope data collection has evolved overtime (Bowen & Vaughan 1973). It used to collect imaging data, which I will talk about first. In recent times, such as the 16th SDSS data release, the Du Pont telescope is used for collecting spectra data (Ahumada et al. 2020), which is what I'll talk about later.

λ (Å)	Angle (arcmin)	f_h (mm)	h/dh (mm)	D (mm)	ϵ (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), Telescope Optical Performance for the Spectrographic Mode: Average Focus

During its optical era, the telescope is a modified Ritchey-Chrétien design with Gascoigne correcton and a 100-inch primary mirror (Bowen & Vaughan 1973). Approximately 40% of the light is reflected to the secondary mirror, resulting in only a 16% loss of light at that stage. (Bowen & Vaughan 1973)

The du Pont Telescope used 18.9 inch nonvignetted plates in order to minimize vignetting (Bowen & Vaughan 1973). 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 (Richards 2020). Because of the nonvignetted plates, the du Pont Telescope experiences an exceptionally low 3% percent loss of light (Bowen & Vaughan 1973).

Another technology the du Pont Telescope applied was a Gascoigne corrector plate. The plate helped with data collection. The Gasciogne corrector plate was abled to be moved, which could help optimize the collection of light in a wanted wavelength (Bowen & Vaughan 1973). Given a seperation of 1000 mm from the end of the corrector plate to the focus gave an image with a minimized astigmatism for a refractive index of n = 1.47 (Bowen & Vaughan 1973). At a given wavelength, the change of length which minimized astigmatism is described in Bowen's paper as

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

Where ΔL is the change in seperation in millimeters and Δn is the difference between a refractive index of 1.47 and the index wanted.

The last technology the du Pont Telescope used was conical baffles. The reason for this was to promote shielding in the telescope (Bowen & Vaughan 1973). As explained in the Bowen paper, shielding was necessary in order to protect the photographic plate from light that escaped from the secondary lense due to long time exposure. the conic baffles were located in the space between the primary and secondary lenses in the plate (Bowen & Vaughan 1973). Theoretically, the conic baffles had the disadvantage of producing a diffraction pattern. However, as explained by Bowen, this did not majorly affect the images of stars (Bowen & Vaughan 1973).

Towards the mid 2010's the du Pont telescope has shifted focus towards spectra data (Ahumada et al. 2020). The du Pont telescope alongside the Foundational telescope used spectrographs during the APOGEE-2 survey and sampled approximately 400,000 stars (Ahumada et al. 2020). The du Pont telescope uses a fiber-optic system consisting of 300 short fibers that are used throughout the night (Ahumada et al. 2020). These fibers can collect observations up to 10 plates per night which are stored on five cartridges (Ahumada et al. 2020). The camera uses a 1024 x 1024 pixel ccd, with the most effective wavelength for the camera being approximately 7600 Å (Ahumada et al. 2020)

(3) The New Mexico State University (NMSU) Telescope: The NMSU telescope uses a camera that has a 2048 x 2048 CCD; the camera is controlled by a linux computer, which is connected by fiber optic cables (Holtzman et al. 2010). The data collection of the NMSU telescope is almost fully automated using C++ (Holtzman et al. 2010). 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. The NMSU telescope was used to observe stars too bright to observe with the larger-aperature telecsopes. It is no longer used in SDSS-V (Collaboration 2025).

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 (Gunn et al. 2006). The data pipeline of the SDSS can be further sub-divided as 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. This subpipeline uses vxWorks to initalize the processing sequence by composing star cutouts and column quartiles collected from the CCD's mentioned before (Lupton et al. 2001)

The second subpipline is the MT pipeline. This pipeline processes the data collected from the Photometric telescope. These data are used to calculate important parameters for the 2.5 m telescope scans, such as extinction and zero-points (Lupton et al. 2001).

The third pipeline is the serial stamp collecing (SSC) pipeline (Lupton et al. 2001). The SSC reorganizes the star cutouts collected from previous pipelines in order to prepare data for the subsequent processing (Lupton et al. 2001).

The Astrometric pipeline follows and estimates the average position of stars using data collected from the Astroline and SSC pipelines. It then converts the pixel coordinates to celestial coordinates (α, δ) (Lupton et al. 2001).

The next stage is the Postage Stamp Pipeline (PSP). The PSP estimates data quality by calculating factors such as the flat field vectors, bias drift, and sky levels (Lupton et al. 2001).

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 image datasets and cataloging the images (Lupton et al. 2001).

The calibration pipeline takes data from the MT and Frames pipeline and converts the counts into calibrated flux densities (Lupton et al. 2001).

(2) Spectroscopy Data Pipeline:

Real-Time Processing—

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

Data Collection and Storage—The SDSS collected around 16 TB of data over a decade in their data release 7 (Juric et al. 2017). Yet the LSST is expected to collect 20 TB of data per night (NSF-DOE Vera C. Rubin Observatory 2025).

The LSST pipeline consists of approximately 750000 in Python and uses relevant libraries such as SciPy [18] and AstroPy [19] (NSF-DOE Vera C. Rubin Observatory 2025). The pipeline also contains approximately 220000 lines of C++ to ensure efficient performance (NSF-DOE Vera C. Rubin Observatory 2025). The tool `pybind11` [20] is needed to parse from Python to C++, and ndarray objects are able to be converted from C++ arrays (NSF-DOE Vera C. Rubin Observatory 2025).

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 (NSF-DOE Vera C. Rubin Observatory 2025).

One instance of a task is the `PipelineTask`, which serves to organize subtasks. These subtasks each have their own purpose (NSF-DOE Vera C. Rubin Observatory 2025). 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 (NSF-DOE Vera C. Rubin Observatory 2025). 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 (NSF-DOE Vera C. Rubin Observatory 2025). The task that first handles processing the catalogued images is the `SourceDetectionTask` (NSF-DOE Vera C. Rubin Observatory 2025). 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 (NSF-DOE Vera C. Rubin Observatory 2025).

Another task that the LSST uses to process data is `MaskStreaksTask` (NSF-DOE Vera C. Rubin Observatory 2025). The task serves to mask pixels from streaks from other satellites. It identifies streaks using a Canny Filter and the Kernel-Based Hough Transform (NSF-DOE Vera C. Rubin Observatory 2025; Fernandes & Oliveira 2008). This task is combined with the deblending of collected images allows for the LSST to accurately identify objects.

[18] https://scipy.org/
[19] https://www.astropy.org/
[20] https://pybind11.readthedocs.io/en/stable/index.html

Real-Time Processing—Due to the sheer volume of data produced by the LSST, processing data in real time is important. It is vital to generate alerts for transient sources in a timely manner for prompt follow-up observations. The Automated Alert Streams to Real-Time Observations (AAS2RTO) is a software system to filter and prioritize alerts (Sedgewick et al. 2025) This may include spectroscopic follow up on the Danish 1.54 m telescope (Sedgewick et al. 2025).

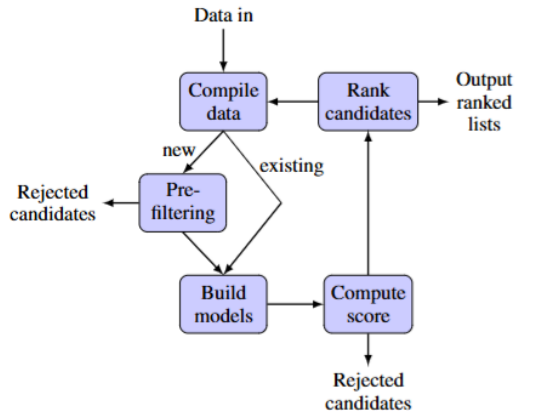

Figure 3. A graph showing the AAS2RTO process to filter and prioritize transient candidates (Figure 2 from Sedgewick et al. 2025)

The first step of AAS2TRO is compiling an unordered set of candidates based on user paremeters, and computing an interest value for each candidate (Sedgewick et al. 2025). AAS2RTO recieves data about new observations from alert brokers such as FINK [21] and Lasair [22] (Sedgewick et al. 2025).

The LSST sends raw data to alert brokers in packets. These packets consist of information about any astrophysical object that underwent a change of brightness or position (Sedgewick et al. 2025). The brokers then filter the data for false positive alerts and send the filtered data to AAS2RTO (Sedgewick et al. 2025).

The next step of AAS2RTO is to prefilter the data (Sedgewick et al. 2025). This is done using a rough scoring function to prefilter bad data effectively before using more costly models (Sedgewick et al. 2025). Then, AAS2RTO fit models in order to further identify good candidates. After that, AAS2RTO computes the scores of candidates to rank observations in descending order and removes rejected candidates (Sedgewick et al. 2025).

An application of AAS2TRO, as described in (Sedgewick et al. 2025), is to classify type Ia supernovae (SNe Ia). The prefilter step uses a specialized model. Combining four functions, described below, and determining whether the score is positive (good data), or negative (bad data) (Sedgewick et al. 2025). The first equation is a magnitude factor x_{mag}: (Sedgewick et al. 2025)

$$x_{\text{mag}} = 10^{0.5\times(18.5-m)} \tag{2}$$

[21] https://fink-broker.org/about/
[22] https://lasair-ztf.lsst.ac.uk/

Where m is the transient magnitude. This equation promotes brighter supernovae over fainter ones. Supernovae fainter than 18.5 magnitude are then removed from the scoring list, but not completely deleted, as they may become brighter (Sedgewick et al. 2025).

The next function is the peak brightness: (Sedgewick et al. 2025)

$$x_{\text{peak}} = A \times \exp\left[-\frac{(t_{\text{obs}} - t_0)^2}{2\sigma^2}\right] \tag{3}$$

This Gaussian function serves to emphasize objects that are near their predicted peak using light-curve models such as the Spectral Adaptive Lightcurve Template (SALT) model (Sedgewick et al. 2025). The term t_{obs} is the observation time of the supernovae (Sedgewick et al. 2025). A is the amplitude parameter (set to $A = 30$), and σ is the peak width and is usually set as $\sigma = 1$ day (Sedgewick et al. 2025). This function prioritizes supernovae close to their peak (Sedgewick et al. 2025). If $|t_{\text{obs}} - t_0| > 4$, then x_{peak} is 10^{-2} (Sedgewick et al. 2025). Lastly, if the SALT model fails, x_{peak} is set to 1.0 (Sedgewick et al. 2025).

The third function is as follows:

$$x^k_{\text{rise}} = \frac{\sum_{i=1}^{N_k - 1} \left[m^k_{i+1} < m^k_i \right]}{N_k - 1} \tag{4}$$

Where k is the photometric band index. m^k_i is the magnitude of the ith detection in the kth band, N_k is the number of detections in the k band (Sedgewick et al. 2025). The numerator of the equation is a boolean expression. This equation is used to quantify if the supernovae is becoming brighter or dimmer using current and earlier observations (Sedgewick et al. 2025).

The last equation is:

$$x_{\text{span}} = \begin{cases} 1, & T < 20 \text{ days} \\ L(T; r, x_m), & \text{otherwise} \end{cases} \tag{5}$$

where L is defined as:

$$L(x; r, x_m) = \frac{1}{1 + \exp(-r(x - x_m))} \tag{6}$$

Here, x_m is the day on which the brightness peaks. The function x_{span} quantifies time since the first observation (Sedgewick et al. 2025). Functions x_{span} and x_{peak} are useful proxies when the SALT model fit fails (Sedgewick et al. 2025). Observations older than 30 days are discarded (Sedgewick et al. 2025)

These four functions combine into the SNe Ia score: (Sedgewick et al. 2025)

$$S_{\text{Ia}} = S_{\text{base}}\, x_{\text{mag}}\, x_{\text{peak}}\, x_{\text{rise}}\, x_{\text{span}} \tag{7}$$

where $S_{base} = 1$ (Sedgewick et al. 2025). Next, the final score function is calculated from S_{Ia} and the object visibility function x_{vis}: (Sedgewick et al. 2025).

$$x_{\text{vis}} = \left(\frac{A_{\text{vis}}}{(a_{\text{ref}} - a_{\text{min}})(t_{\text{SR}} - t_{\text{obs}})} \right)^{-1} \tag{8}$$

where A_{vis} is:

$$A_{\rm vis} = \int_{t_{\rm obs}}^{t_{\rm SR}} [a(t) - a_{\rm min}]\, dt \tag{9}$$

Here $a_{ref} = 90°$ is the reference altitude used for normalization, $a(t)$ is the altitude at a given time, a_{min} is the minimum observable altitude, t_{SR} is the expected time of the next sunrise, t_{obs} is the time of the observation (Sedgewick et al. 2025). Functions x_{vis} weights objects by the intensity of the observation. The final score, S_{DK154}, is calculated as:

$$S_{DK154} = S_{\rm Ia}\, x_{vis} \tag{10}$$

AAS2RTO then ranks the observations in descending priority and removes negative scores for rejected candidates.

AAS2RTO is only one of many programs used to deal with real-time processing for the LSST. Such programs demonstrate the challenges of big data in astronomy.

2.2. *The Radio Big Data Pipeline*

2.2.1. *The MeerKAT*

Data Processing—The MeerKAT data processing pipeline is split into three parts, the calibration pipeline, the continuum imaging pipeline, and the spectral imaging pipeline (Ratcliffe 2021).

(1) The Calibration Pipeline: The pipeline starts by choosing an antenna as a reference (Ratcliffe 2021). The first scans for a given antenna are then averaged and the fourier transform is applied. The peak-to-noise ratio is the ratio of the data maximum to the peak rms noise (Ratcliffe 2021). The antenna with the highest median peak-to-noise ratio over every baseline phase is chosen as the reference antenna (Ratcliffe 2021). This antenna will have all of its phase calibration solutions set to zero (Ratcliffe 2021).

The purpose of the calibration pipeline is to ensure that instrumental antenna errors are calibrated. At the start of every observation, the reference antenna is evaluated based on certain flags, such as data loss (Ratcliffe 2021). If 80 percent or more of the data is flagged, a new reference antenna is determined using the aforementioned process (Ratcliffe 2021).

(2) The Continuum Pipeline: The second pipeline is the continuum pipeline. This pipeline produces continuum images using the OBIT software package (Ratcliffe 2021; Cotton 2008). The OBIT package reads the UV data. The OBIT `MFImage` task is used to perform wide-band, wide-field imaging (Ratcliffe 2021; Cotton & Schwab 2010). Once the data are read, they are split into scans, which are then averaged and combined into a dataset.

This dataset is then split into eight 107 MHz intermediate frequencies (Ratcliffe 2021). The UV data's headers contain the number of antennas, the number of baselines, the number of channels, and the number of polarization (Ratcliffe 2021). In the 48-antenna Meerkat's array, short baselines dominate (Ratcliffe 2021). This allows baseline-dependent averaging to reduce data volume (Ratcliffe 2021). This step reduces the data volume by about 3-4 times (Ratcliffe 2021).

The `MFImage` task uses joint frequency deconvolution to handle wide-band effects in wide-band images (Ratcliffe 2021). Normally, Meerkat uses approximately 140 circular image facets with a size

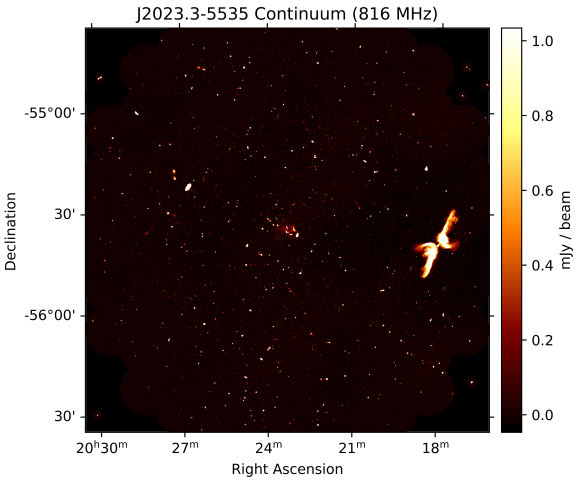

Figure 4. An example UHF continuum thumbnail image produced automatically by the continuum imaging pipeline is shown above, (Ratcliffe 2021)

of 6 arcminutes to cover 1 degree from the phase center (Ratcliffe 2021). Additionally, there are facets of 1 degree that use the SUMSS or NVSS catalogue to cover a radius of 2.5 degrees from the phase center (Ratcliffe 2021). These 1-degree facets are used for phenomena that are ancipiated to have a flux density greater than 5 mJy (Ratcliffe 2021). The facets integrate wide-band imaging effects. The frequency band is split into 10 components (Ratcliffe 2021). During joint-frequency deconvolution, the brightest sources in the dataset are found and subtracted from the slices of data individualy using the CLEAN algorithm (Ratcliffe 2021).

Self-calibration is then performed in two rounds. The first round uses CLEAN components with a flux density above 1 mJy. This yields approximately 1000 CLEAN components (Ratcliffe 2021). The data are then self calibrated in a second round, using a CLEAN threshhold of 100 μJy. The second round yields approximately 10000 CLEAN components (Ratcliffe 2021). After this process, the field continuum images are produced. Lastly, the self calibrated data is converted into AIPS format (Ratcliffe 2021). The sky model, made up of the CLEAN components, is stored in AIPS CC format. The flux density of all the 10 frequency components are summed (Ratcliffe 2021). The merged flux density is then fitted with a second degree polynomial over frequency and subtracted from the image (Ratcliffe 2021). Finally, the fully processed images are converted into FITS and PNG files and archived (Ratcliffe 2021).

(3) The Spectral Pipeline: The purpose of the spectral imaging pipeline is to produce high-quality spectral line images effectively (Ratcliffe 2021). Spectral channels are independent of each other, and can be processed in parallel (Ratcliffe 2021). However, the raw resulting data are recieved in time-major order, and this data structure must be transposed in channel-major order (Ratcliffe 2021). In order to solve this issue, a visibility writer is used. The writer stores the visibilities on a Ceph cluster [23] with chunks each across 64 spectral channels (Ratcliffe 2021). This choice in the chunk size is not optimized, but it does avoid issues with RAM and memory allocation (Ratcliffe 2021).

[23] https://ceph.io/en/

Using CUDA [24], every channel then is imaged separately. The use of CUDA allows for the processing to speed up due to the usage of NVIDIA GPUs (Ratcliffe 2021). The iteration over W slice in each chanel is shown in psuedo-code in Fig 4 (Fig. 1 of (Ratcliffe 2021))

1. For each W slice[1]
    a. Apply image-plane W term and taper correction to the model image
    b. FFT the result to get a UV grid.
    c. For each batch of visibilities
        i. Predict visibilities by degridding, and subtract them from the measured visibilities in place.
        ii. Grid the resulting batch of visibilities.
    d. Inverse FFT the grid.
    e. Apply image-plane W term and taper correction in the image plane.
    f. Add the result to the dirty image.
2. Apply CLEAN, adding new components to the model image.

Figure 5. The inner cyclic algorithm used by the MeerKAT spectral line pipeline, (Ratcliffe 2021)

The chunk system only partially solves the problem of transposing the data and further steps are necessary (Ratcliffe 2021). The visibilities in each chunk are ordered by channel, w slice (wide-field imaging plane), and baseline. The data chunks inherently store measurements such as UVW coordinateas and parallatic angles (Ratcliffe 2021). At this point, conservative baseline-depedent averaging is also applied to the visibilities (Ratcliffe 2021). 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 (Ratcliffe 2021).

2.2.2. *The Square Kilometre Array (SKA)*

3. RESULTS

3.1. *The Optical Data Results*

3.1.1. *The Sloan Digital Sky Survey (SDSS)*

Since 1998, the SDSS has published nineteen data releases over five project generations, each with their own goals. Most recently, SDSS-IV used spectroscopic surveys to detect cosmological objects (Ahumada et al. 2020). The sixteenth data release (DR16) is the first one for SDSS-IV (Ahumada et al. 2020). While the seventeenth data release (DR17) will be the lass for SDSS-IV. I plan to compare these two data releases. DR18 is ommited because it was mainly used to set the foundation for future SDSS-V data releases by introducing new models, functions, strategies, and more (Almeida et al. 2023). The nineteenth release (DR19) is also excluded because it is only a preview of what is achievable by the SDSS-V (Collaboration et al. 2025) The two data releases compared quantify the increase in data volume within the same generation.

Every successive data release has grown in terms of volume of data (Fig 6). The first few generations increased only slightly, and the later generations increased exponentially. This is a clear example of

[24] https://developer.nvidia.com/cuda-zone

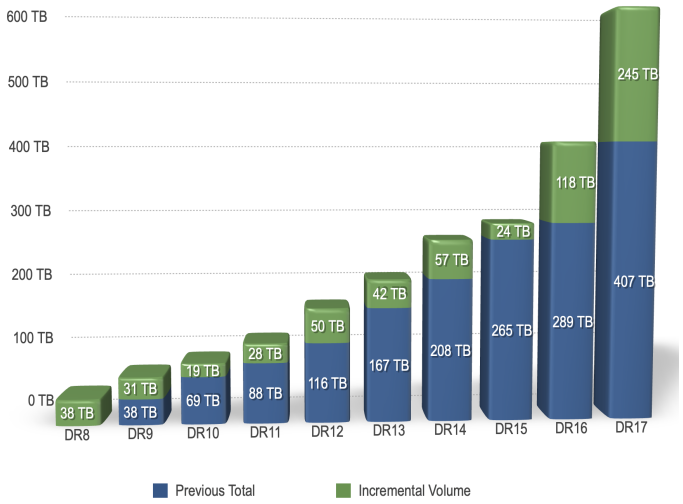

Figure 6. The increase of data volume over multiple SDSS data releases (Figure from the SDSS website[a])

[a] https://www.sdss4.org/dr17/data_access/volume/

Moore's Law. DR16 accounts for approximately 18.1% of the total SDSS-IV data, and DR17 accounts for approximately 37.7% of the SDSS-IV's total data. The seventeenth data release consists of over 46 million new files, which is the reason for the massive increase in data volume (Almeida et al. 2023). This increase in data can be summarized in four reasons (Almeida et al. 2023). The first reason is because DR17 includes the entirety of the APOGEE-2 survey, which combined an additional 879,437 infrared spectral measurements (Almeida et al. 2023). The second reason is because the MaNGA survey also completed at this time (Almeida et al. 2023). The MaNGA survey release included 11273 cubes compared to DR16's 4824 cubes (Almeida et al. 2023). These cubes hold 3D data, which two spacial dimensions and one wavelength dimension (SDSS 2019). The third reason is because of the eBOSS survey (Almeida et al. 2023). Also, DR17 contains 25 value-added catalogues (VAC) that were either updates, or new (Almeida et al. 2023). These VACs account for a substantial portion of the total data in DR17 (Almeida et al. 2023). The final reason is that DR17 includes all the previous SDSS data releases (Almeida et al. 2023). The data were reprocessed with the newest pipelines at that time (Almeida et al. 2023). The updated pipelines generated more data (Almeida et al. 2023)

The comparison of only two consecutive releases has shown a massive increase in data volume. This trend is likely to explode with SDSS-V, which we will discuss later.

New text starts here:

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

LSST has yet to release data, but it is expected to observe an 18,000 deg^2 approximately 800 times during a 10 year span (Ivezić et al. 2019). These 800 iterations will result in a total of 32 trillion observations (Ivezić et al. 2019). LSST is estimated to produce around 20 terabytes of data per night (LSST ????), resulting in a total of approximately 73 petabytes of raw data over the 10 years. This would make the total estimated data volume of the LSST approximately 111.96 times bigger than the total data volume of the SDSS pre-DR18.

Another key feature promised of the LSST is the LSST Science Platform (LSP) (Jurić et al. 2019). LSP is a collection of web-based tools and services (Jurić et al. 2019). LSP aims to centralize the access and analysis of LSST data for the scientific community (Jurić et al. 2019). LSP is organized into three sections, refered to as Aspects (Jurić et al. 2019). The first aspect is the Portal Aspect (Jurić et al. 2019). The portal aspect is a website that allows users to analyze and visualize data from both past archives, such as the Infrared Science Archive, the SDSS archive, etc, and LSST images (Jurić et al. 2019). The Portal Aspect allows users to download data, make plots, and mask data to find specific data points (Jurić et al. 2019).

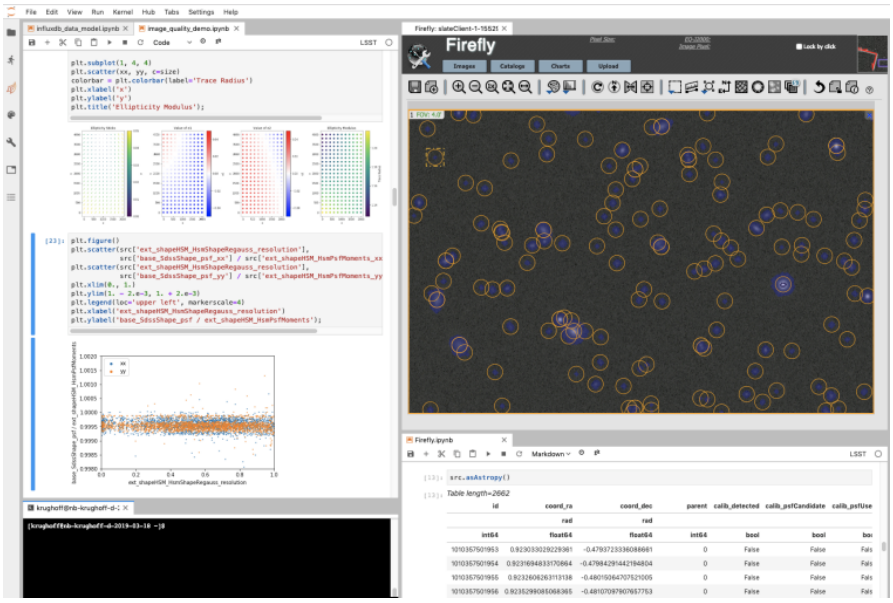

Figure 7. An image of the Notebook Aspect's UI (Figure 4 from Jurić et al. (2019))

The second aspect of the LSP is the Notebook Aspect (Jurić et al. 2019). Compared to the first aspect, the Notebook Aspect allows users to handle more detailed data analysis through the use of Jupyter notebooks (Jurić et al. 2019). Computations in said notebooks will be done using resources from the LSST Data Access center, and not user-local resources (Jurić et al. 2019). This is done because storing and analyising the large data will hinder users without resources (Jurić et al. 2019). The notebooks will come preinstalled with important libraries, such as AstroPy, the LSST science pipelines, visualization libraries, and more (Jurić et al. 2019). These notebooks will be stored as user-generated data products in the LSST Data Access Center (Jurić et al. 2019).

The final aspect of the LSP is the Web API Aspect (Jurić et al. 2019).The Web API aspect allows users to extract data using APIs (Jurić et al. 2019). Instead of working with notebooks, users can query the data to use for extraneous astronomical tools (Jurić et al. 2019). Another notable feature of the LSP is the use of collaboration (Jurić et al. 2019). The LSP allows users to share and edit files and catalogues that can be used for all three aspects (Jurić et al. 2019).

Note: Will do this later. Seems super complex and I think it honestly might deserve its own attention

Another result of the LSST's development is the parallel construction of Machine learning and AI tools. An example of such tools is a spatio-temporal engine which uses three AI models to observe

supernovae (Kodi Ramanah et al. 2022). The engine ran simulations of LSST data in which it reached accuracy of around 99% (Kodi Ramanah et al. 2022).

The first model is the single-epoch model (Kodi Ramanah et al. 2022). The single-epoch model uses a convolutional neural network (CNN) (Kodi Ramanah et al. 2022). The CNN is made up of two consecutive convolutional layers followed by two consecutive max-pooling layers, this repeats for a total of three times (Kodi Ramanah et al. 2022). After the data is processed, it is flattened into a 1D vector, where it is processed through two fully connected layers (Kodi Ramanah et al. 2022). The latter layer converts all the pixels into a probability score from [0,1] (Kodi Ramanah et al. 2022)

The second model is the multi-epoch model (Kodi Ramanah et al. 2022).

The final model is the spatio-temporal mofel (Kodi Ramanah et al. 2022).

3.2. *The Radio Data Results*

3.2.1. *The MeerKAT*

Note: Will work on this next. It's hard to find numbers for the MeerKat and SKA

3.2.2. *The Square Kilometre Array (SKA)*

The SKA will collect over 700 petabytes of data per year (SKAO 2025).

New text ends here:

4. DISCUSSION

4.1. *Open Source Policies and Transparency*

4.1.1. *The SDSS Policy*

The SDSS-IV collaboration states that all of its software is open source under the BSD-3 license. However, the SDSS outlines practices for users who wish to reuse or extend the SDSS software. Most importantly, proper citation of software and websites is required.

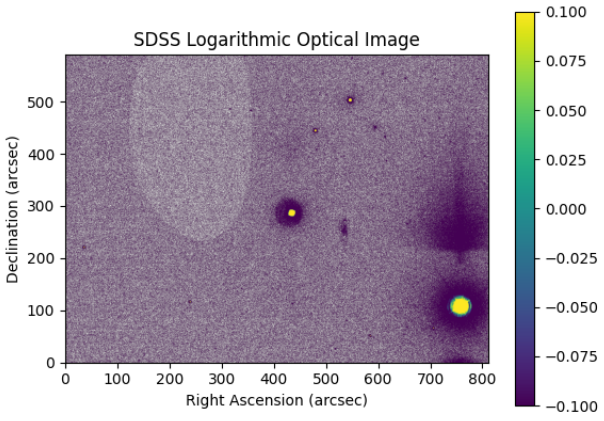

Figure 8. A spectrographical image obtained using data collected from SDSS

The SDSS has also implemented digital object identifiers (DOI) in all software code. These DOIs allow software and data to be easily identified, which is important for ownership. The SDSS team also promotes transparency in coding by Git and SVN [25] to version code repositories.

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*

Despite providing fully public data, the LSST has committed to open source software and a transparent pipeline. The LSST has declared that any source code made for data management must have an Open Source Initiative license (Observatory ????). Open source data-management software must be issued under the GNU Public license (Observatory ????). Otherwise, users are free to choose any Open Source initative approved license (Observatory ????).

4.1.3. *The MeerKAT Policy*

The South African Radio Astronomy Observatory (SARAO) has expressed that MeerKAT data products must be open source with adequate acknowledgement through two steps (Camilo 2024). First and foremost, all MeerKAT data used must come from the MeerKAT ADS Library[26] (Camilo 2024). Secondly, any publication that contains MeerKAT data, the author must state the following statement: "The MeerKAT telescope is operated by the South African Radio Astronomy Observatory, which is a facility of the National Research Foundation, an agency of the Department of Science and Innovation." (Camilo 2024).

Data products can be released into the MeerKAT archive by contacting the SARAO. (Camilo 2024). The SARAO has developed a help desk [27] from which users can ask general questions (Camilo 2024). If users have questions on policies of the MeerKAT, they are advised to communicate them to the SARAO chief scientist (Camilo 2024). There is also a SARAO Users Committee, reporting to the SARAO MD (Camilo 2024).

New text starts here:

4.1.4. *The SKA Policy*

The SKA has explained that all data collected is owned by the SKAO (Ball 2023). The data will be made open source after a time period as made by the SKAO council (Ball 2023)

4.1.5. *The Importance of Open Source Data in Astronomy*

The field of Astronomy is inherently data-driven, so open source data policies have become central to Astronomy. The SDSS, LSST, and MeerKat have all taken a stance to make data open source as soon as it is obtained. The SKA is the only project that restricts the use of publizing data for a given time period. I believe having open source data and software is crucial for many reasons.

One reason is because it allows for data to be accessible for those without necessary resources. Students and scientists who do not have the necessary resources may rely on the public data in order

[25] https://subversion.apache.org/
[26] https://ui.adsabs.harvard.edu/public-libraries/wmc9yO6IQ3mUZCPx7MQRxg
[27] https://commons.datacite.org/doi.org?query=client.uid%3Awhno.ljncxe&resource-type=dataset

to support a claim. In essense, restricting data in any sense is immoral because everyone has a right to knowledge.

Another reason open source data is important is because it allows for the advancement of science. With the data being public, more people will analyze the data and new discoveries may be found. Public data will also in turn bring more attention to the survey, so users may build new, more advanced, software for the survey.

A third reason that open source data is important is because it confirms the reliability of claims. If a scientist makes a claim based on results made from data, others must reproduce the same result to confirm the claim. So if the data wasn't public, only scientists within the same team would be able to reproduce the same results. These scientists would more than likely support the claim, which creates a conflict of interest and hinders the public from the truth.

Similar to the previous reason, having a transparent pipeline is also important. This is because it builds public trust for the survey. If users are allowed to see every step of the data processing, they are able to understand how the data was transformed, rather than blindly being handed it. It also allows for users to point out hidden biases within the pipeline that may skew the data.

4.2. *The Increase of Skill Needed*

As explained by Moore's Law, the exponential increase of data collected has led to an exponential increase in complexity of handling the data. In the past, astronomy has emphasized collecting data through observations. On the other hand, modern astronomy relies on high-power computing, software engineering, and data analysis. Recent surveys such LSST and SKA produce volumes of data that would be impossible to process in a lifetime. The data is processed automatically through multiple pipelines written in Python and C++ code that only increase in complexity over time. As a result of this, more skill is needed from scientists. This may be crucial especially for new scientists. They did not learn as the surveys became more complex, so they have to work from the ground up when compared to experienced scientists.

Despite the growing complexity, the evolution of skills required may also create opportunities not traditional to astronomy. What was once a field mainly for astronomers, now concerns astronomers, engineers, computer scientists, statisticians, etc. This This increase of collaboration may encourage innovation as these scientists have different skillsets and mindsets. While the increase of skill needed has made observational astronomy more difficult, it has allowed for more collaboration and innovation.

4.3. *The Storage and Sustainability of the Data*

In the past, the storage of data collected was an afterthought. The main focus of collecting data was drawing claims from them. Due to the expansion of data into the petabyte scales, the handling of the data has become more challenging. Challenges arise, such where to store the data. Data must be stored in data centers with high capacity. These data centers must also have enough bandwidth to allow users to retrieve data remotely. As a result of this, more money must be allocated to store the data, which may take away from funds that could be used to further develop instruments.

Another concern is the sustainability of the data. This is currently not a glaring issue, but as data volume increases, so too does the energy needed to store it in the data centers and process the data. Especially due to the rise of AI in astronomy, the energy needed may become significant enough to negatively impact the enviroment.

4.4. *The Rise of AI/ML in Surveys*

4.5. *The Future of Big Data in Astronomy*

Not done yet with this

From being completely handled by humans, to having minimal human interaction, the processing of data in astronomy has evolved overtime. In recent times, it has evolved at a more rapid pace. In the present, astronomy has entered an age of unrivaled data volume and complexity. Through the use of observing the four main surveys we have observed, it is clear that this behavior will only continue in the future.

The evolution of data has also caused a shift in priorities, which cause problems to arise. As seen through software utilitizing new technology such as machine learning and AI, I believe that the collecting of data will not be an issue in the future. By looking at the open source policies of all surveys, I believe that open data and transparency will also not be an issue in the future. I think the real issue is the sustainability of the data and the need for more skill.

New text ends here:

5. CONCLUSION

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

Ahumada, R., Prieto, C. A., Almeida, A., et al. 2020, The Astrophysical Journal Supplement Series, 249, 3, doi: 10.3847/1538-4365/ab929e

Almeida, A., Anderson, S. F., Argudo-Fernández, M., et al. 2023, The Astrophysical Journal Supplement Series, 267, 44, doi: 10.3847/1538-4365/acda98
Ball, L. 2023
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
Collaboration, S. D. S. S. 2025, Instruments
Collaboration, SDSS., Pallathadka, G. A., Aghakhanloo, M., et al. 2025, The Nineteenth Data Release of the Sloan Digital Sky Survey, arXiv, doi: 10.48550/arXiv.2507.07093
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
Jurić, M., Ciardi, D. R., Dubois-Felsmann, G. P., & Guy, L. P. 2019, LSE-319: LSST Science Platform Vision Document, NSF-DOE Vera C. Rubin Observatory, doi: 10.71929/RUBIN/2587242
Juric, M., Kantor, J., Lim, K.-T., et al. 2017
Kodi Ramanah, D., Arendse, N., & Wojtak, R. 2022, Monthly Notices of the Royal Astronomical Society, 512, 5404, doi: 10.1093/mnras/stac838
Lesser, M. 2015, Publications of the Astronomical Society of the Pacific, 127, 1097, doi: 10.1086/684054
LSST. ????, Data Management, https://www.lsst.org/about/dm
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
Observatory, V. R. ????, LSST Licensing Overview, https://developer.lsst.io/legal/licensing-overview.html
Ratcliffe, S. 2021, SDP Pipelines Overview, https://skaafrica.atlassian.net/wiki/spaces/ESDKB/pages/
Richards, S. 2020, What Is Vignetting?

SDSS. 2019, MaNGA Data, https://www.sdss4.org/dr17/manga/manga-data/

Sedgewick, A., Gall, C., Izzo, L., et al. 2025, Astronomy & Astrophysics, 698, A153, doi: 10.1051/0004-6361/202452099

SKAO. 2025, Handling a Deluge of Big Data, https://www.skao.int/en/explore/big-data

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