Draft version September 28, 2025
Typeset using LaTeX **preprint2** style in AASTeX631

The New Age of Big Data In Astronomy: A Review of on the SKA & LSST

Mathew Icho
The University of Illinois at Urbana-Champaign

ABSTRACT

Write my abstract here

Contents

1. INTRODUCTION

The concept of data has long been a principal concern 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, is empirical evidence. Empirical evidence refers to data collected through traditional means, such as direct observation or experimentation of natural phenomena using sensory perception or basic instruments. The primary purpose of empirical evidence is to identify patterns that allow scientists to develop a fundamental understanding of the natural world. Throughout much acquisition human history, empirical evidence represented the most prominent method for extracting knowledge from nature. 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 primarily collected and cataloged data on the position of astronomical bodies using naked-eye observations. However, this method of data collection is associated with several limitations. 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 represents the second most prominent mode of scientific inquiry in terms of longevity. The primary purpose of analytical evidence is to construct mathematical formulas and theoretical frameworks based on empirical data. Unlike the first paradigm, 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, a student of Tycho Brahe, 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, a relatively recent development with respect to the first two. The purpose of simulation evidence is to model natural phenomena that are too complex to solve analytically by hand. Its central role is to enable interpolation and extrapolation of data using computational techniques grounded in known physical laws. In astronomy, an example is the use of N-body simulations to study the dynamical evolution of planetary systems and galaxies. By simulating the gravitational interactions between multiple bodies simultaneously, astronomers can decipher theoretical structures and determine long-term behaviors that would be analytically impossible to solve.

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 one of the most prominent trailblazers of this paradigm. Modern observatories now generate petabyte-scale data that need new strategies for data management and analysis. The fourth paradigm in turn reshapes the scientific process itself. Instead of discoveries being made from observation or theory, they are now being made from interpreting massive data sets. However, these advances also expose alarming issues, including bottlenecks in the data pipeline, the storage of aforementioned data, the increase in skill needed to handle the data, and concerns regarding open access to data. The field of astronomy is both a beneficiary and a victim of this data-intensive transition.

The exponential growth of data acquisition can be attributed to Moore's law. Moore's law is an observation coined by Cofounder of Intel, Gordon E. Moore in his paper titled, "Cramming more components onto integrated circuits". In said paper, Moore explains that the number of components that make up an integrated circuit increase approximately at a rate of a factor of 2 per year. Moore also stated that this growth is not sustainable more many reasons, the most relevant being the fact that techniques to handle such complex circuits lag behind in terms of development Moore (2006).

Moore's law can be seen in many data-intensive fields, such as astronomy. When applied, it explains both the recent development of Big Data in astronomy, and accurately predicts the present issue that methods being used for analyzing said data is lagging behind, causing the mentioned issues.

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, The Sloan Digital Sky Survey (SDSS), The Legacy Survey of Space and Time (LSST), and The Square Kilometre Array (SKA). These facilities collectively highlight the scope of contemporary astronomical data, the methods of its acquisition, their relative successes, the ongoing challenges, and the solutions currently in use. I plan on doing this by reviewing the approach that all four case studies have taken or plan to take to collect data.

1.3. *A Qualitative Analysis of Relevant Survey Data*

To demonstrate the rise of Big Data in Astronomy, we must first examine the components that make up the SDSS. The SDSS is vital to this paper, as it is one of the earliest large-scale optical surveys that signifies the start of the fourth paradigm. Because of its relative early involement in the Big Data stage of astronomy, and its use of collecting optical data, I plan to compare the SDSS to LSST. The SDSS consists of three main telescopes.

The first of the three is "The Sloan Foundation 2.5m Telescope". The Sloan Foundational Telescope is stationed at the Apache Point Observatory in New Mexico, where it observes the sky in the northern hemisphere. The Sloan foundational telescope is able to observe a 3° field of view through the use of its two corrector lenses, that help with distortion. Gunn et al. (2006)

Another vital telescope used in the SDSS project is "The Irénée du Pont Telescope at Las Campanas Observatory" . The Irénée du Pont Telescope differs from the 2.5m Telescope as it is stationed in Chile, where it observes the southern hemisphere instead. Similar to the first mentioned telescope, the Irénée du Pont Telescope displays a 2.1° field of view with only one corrector lens. Bowen & Vaughan (1973)

The third yet most vital telescope is the "NMSU 1-meter Telescope". The NMSU telescope is stationed in the Apache Point Observatory alongside the Sloan Foundational Telescope. The NMSU telescope serves a purpose the former two don't because "Obtaining spectra of these bright sources is a challenge for the Sloan 2.5 m telescope and not practical through drilling and observing specialized plug-plates" Majewski et al. (2017). In essense, by using optical fibers connected to a spectrograph, the NMSU telescope observes stars that are too bright for the other two to observe. The combination of these telescopes allow for both optical data to be collected through multiple surveys Holtzman et al. (2010)

the SDSS is made up of multiple subsurveys. The eBoss, a continuation of BOSS, utilize spectrographs to observe light at a wavelength range of 3600-10,400 ÅDawson et al. (2016). An additional subsurvey is the APOGEE-2, a continuation of APOGEE. It uses spectrographs similar to eBOSS, but APOGEE-2 collectes near-infrared objects Majewski et al. (2017). MaNGA is a subsurvey that collect integral field unit measurements of 10,000 nearby galaxies using spectrographs Bundy et al. (2014a). MARVELS is another subservey that makes up the SDSS, 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 is another survey essential to demonstrate how Big Data has evolved in the field of Astronomy. MeerKAT became fully operational in 2018 in the Northern Cape Province of South Africa. It serves as a precursor to the Square Kilometre Array (SKA), as both facilities focus on the collection of radio data Jonas & the MeerKAT Team (2018). 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 Goedhart (2025).

MeerKAT has conducted and continues to conduct ten major survey projects. For the sake of conciseness, this discussion will focus on five of these surveys. One is the "Looking At the Distant Universe with the MeerKAT Array" (LADUMA) survey. The objective of the LADUMA survey is to "use HI observations to study galaxy evolution over two thirds of the age of the universe" Blyth et al. (2018). LADUMA has used "shorter observations with MeerKAT's "Phase 1" (0.9-1.75 GHz) receivers would be followed by longer observations in an expanded "Phase 4" (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 (MALS) is a survey conducted using MeerKAT using by collecting data about HI and OH absorvers at $z < 0.4$ and $z < 0.7$, where z is the redshift of a galaxy. The reason for using the HI observation specifically is because it is a descriptive tracer of the cold neutral medium in a galaxy. The cold neutral medium give scientist details on what the physical conditions of the interstellar medium of said galaxy is. This, in turn, allows scientists to extrapolate data on the star formation in the galaxy Gupta et al. (2021).

Another survey, the "The Hunt for Dynamic and Explosive Radio Transients with MeerKAT" (ThunderKAT), aims to find, identify and understand high-energy astrophysical processes via their radio emission (often in concert with observations at other wavelengths)." In essence, ThunderKAT analyzes radio data to catalogue high-energy phenomena, including supernovae, microquasars, and similar events Woudt et al. (2018).

Another notable survey under MeerKAT is the MeerKAT HI Observations of Nearby Galactic Objects: Observing Southern Emitters (MHONGOOSE). this survey aims to catalogue the properties of HI gas in "around 30 nearby star-forming spiral and dwarf galaxies to extremely low H i column densities". MHONGOOSE is remarkable by its notably higher sensitivity compared to previous surveys such as HALOGAS and THINGS. 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 survey considered here is the "MeerKAT International GHz Tiered Extragalactic Exploration" (MIGHTEE). MIGHTEE utilizes radio data spanning 900–1670 MHz, achieving a resolution of approximately 6 arcseconds.

Together, these five surveys, along with the remaining projects, emphasize the pivotal role of MeerKAT in the era of Big Data astronomy. They are at the forefront of scientific research and are producing data volumes on the order of petabytes.

Having examined MeerKAT, its only natural to examine its successor, the SKA. The SKA has built on technical and scientific achievements paved by MeerKAT. The SKA covers an area of approximately 3,000 km with antennas, which collect area up to 106 m^2 of the sky. Because of its astonishing techniqical prowess and complexity, The SKA represents the start of a new frontier for Big Data astronomy. One technology being used is aperture synthesis, which allows for the signals from antennas to be in phase, allowing to reduce noise. Another innovation is the use of large centimeter wavelength antennas, which allow for data to travel across distances at high speeds. This, in turn, allows for data to be analyzed and processed quicker and optimized, which will be discussed in further detail in the Methods section. Dewdney et al. (2009).

Alongside the SKA is its optical counterpart, the LSST, which similarly represents a major advance in the evolution of Big Data in astronomy The LSST is a successor to the SDSS, as both projects observe optical data. The LSST, however, has much more sophisticated goals. The LSST plans to address 4 key scientific issues: Investigating dark energy and dark matter, cataloguing the solar system, collecting data for sky surveys, and mapping the Milky way. To do all this, the LSST uses a 3.2-gigapixel camera with a sampling of 9.6 deg^2 field of view. These cameras are equipped with highly resistant sensors reinforced with silicon Ivezić et al. (2019). All these inventions alongside others allow for the LSST to acquire complex and immense data sets never seen before.

Truly, the SDSS, MeerKAT, SKA, and LSST are the pinnacle of human ingenuiety. They will allow for unprecedented data rates for complex astrophysical events and phenomena. In the following Methods section, I describe how the data is collected, processed, analyzed, and stored. I then compare SDSS and MeerKAT to their respective successor in order to address the issue of inflation of Big Data in astronomy.

2. METHODS

REFERENCES

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

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

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

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

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