Draft version September 26, 2025
Typeset using LaTeX **manuscript** 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 central concern throughout the history of astronomy. Data enables scientists to discern the underlying principles governing natural phenomena, exert control over events, and make reliable predictions. It has played a critical role in other time-sensitive domains 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, from medicine to astronomy, 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 of 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, a representative example is the use of N-body simulations to study the dynamical evolution of planetary systems and galaxies. By simulating the gravitational interactions of many bodies simultaneously, astronomers can investigate emergent structures and long-term behaviors that would be analytically intractable.

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 Main Idea*

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. This paper therefore seeks to review the technical and scientific issues surrounding big data in astronomy 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 astronomical data, the methods of its acquisition, their relative successes, the ongoing challenges, and the solutions currently in use.

What actually are these four astronomical facilities and why did I chose them to represent the inflation of data science in Astronomy? I plan to answer this question in this section.

The reason for observing these four specific astronomical facilities is simple. Although they are inrefutably part of the fourth paradigm, MeerKAT and SDSS are precursors to SKA and LSST.

Because of this, I plan on mainly using SKA and LSST

2. METHODS

REFERENCES

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