

Astronomy in the Era of Big Data: The Status Quo, Challenges, and Future Opportunities

Min Jun Kim
(Qingdao Daewon School)

Kim, Min Jun. 2022. Astronomy in the Era of Big Data: Status Quo, Challenges and Future Opportunities. *Creative Critical Thinking* 9(1), 1-15. The proliferation of astronomical data combined with recent advances in big data computing technology has opened a new horizon in the field of astronomy and astrophysics. In order to objectively evaluate future opportunities in the area of astroinformatics and astrostatistics and to suggest meaningful strategic plans, this paper aims to analyze the status quo and challenges the current astronomical society is facing. By using literature review as the major means of research methodology, it is believed that this paper can offer the audience with as comprehensive understanding of global trends in astroinformatics and astrostatistics. (Qingdao Daewon School)

Keywords big data, astrophysics, astronomy, astronomical database, Astroinformatics, Astrostatistics, computational algorithms, machine learning, virtual observatory

1. Background

In early 2018, astronomers discovered evidence that thousands of black holes might exist near the center of the Milky Way Galaxy. This fascinating finding was not from a new observation using telescopes but from an analysis using long-archived data from NASA's Chandra X-ray Observatory. Until a few decades ago, astronomers had to physically locate themselves at home institutions where they heavily relied on visual observations using optical telescopes. However, with technological developments which have opened a new horizon in the field of astronomical data collection and storage, astronomers can now draw significantly meaningful conclusions by analyzing collected as-

tronomic data while sitting in their labs. Thus, the importance and application of big data is growing in the fields of astronomy and astrophysics. This has resulted in the opportunity to forge new cross-disciplinary fields such as astroinformatics and astrostatistics.

The proliferation of astronomical data combined with recent advances in big data computing technology has enabled a petabyte-scale data processing capability. With large observatory projects currently under construction such as Vera C. Rubin Observatory (LSST) and Square Kilometre Array (SKA), expected to open a digital window for new discoveries, the data volume is expected to increase geometrically which will require even more advanced big data processing capabilities.

Table 1. Ongoing large astronomy observatory projects

Project Title	Location	Timeline	Task
Vera C. Rubin Observatory (LSST)	Cerro Pachon, Northern Chile	Built: 2015-2021 First light: Expected in 2022/2023	Astronomical survey (Large Survey of Space and Time)
Square Kilometre Array (SKA)	Australia, South Africa	Built: 2021-2030 First light: Expected in 2027	World's largest radio telescope array

In this petascale data regime, the next question that needs to be posed is, who has access to this data. Different from many study fields of study where data access rights are highly privileged, most of the sky survey results are made open to the public, allowing everyone across the globe to be able to access these data and freely analyze them. For instance, via total 16 rounds of data release cadence, Sloan Digital Sky Survey (SDSS) results including images, optical spectra, infrared spectra, IFU spectra, stellar library spectra, and catalog data are currently available on the SDSS website.

2. Objectives and research methodology

Given the up-and-coming opportunity big data analytics can offer to the field of astronomy and astrophysics, it is important to fully understand the status quo and challenges of modern astroinformatics, so that we can objectively evaluate future opportunities and suggest strategic plans. As thus, under the title *Astronomy in the Era of Big Data: Status Quo, Challenges, and Future Opportunities*, this paper explores the following three research questions: 1. What is the global status of the utilization of astronomical big data? 2. What are the major challenges faced by the astroinformatics society? and 3. What are the necessary future steps needed for future developments?

The primary research methodology utilized throughout this paper is literature review. First, various project publications and research papers have been searched and screened based on the purpose of the study. Afterwards, each piece of literature that has passed the initial screening has been appraised for its quality and objectivity, followed by information and data compilation and analysis.

3. Significance

The significance of this paper is that it can offer a comprehensive understanding of global trends in astroinformatics and astrostatistics, the two areas of interdisciplinary study which are becoming the dominant source of new astronomical findings. It is believed that this paper can arouse interest for those who are seeking the limitless possibility big data can offer in the era of space travel.

4. The Status Quo: Global space agencies' initiatives

4.1 North America (NASA, CADC)

National Aeronautics and Space Administration (NASA), one of the biggest generators of data in the world, is generating approxi-

mately 12.1TB every day on average. In 2018 IT Talk, NASA has explicitly mentioned that “data and information are cornerstones of NASA’s past, present, and future missions and discoveries,” and that in order to achieve its strategic goal of leveraging data to amplify the benefits of NASA’s mission, NASA will “establish Agency-wide data standards supported by secure, integrated data search and analytics capabilities to accelerate insights and improve the quality of decision making across its complex data-rich environment” (NASA, 2018). Accordingly, NASA is holding various critical projects that archive and analyze big data as can be seen in Table 2.

Table 2. NASA’s big data projects

Project Title	Objectives
Quantum Artificial Intelligence Lab (QuAIL)	Develop quantum computing and algorithms that may address complicated machine learning problems NASA has for its aeronautics and space exploration missions
High-End Computing Capability (HECC)	Perform large-scale modeling, simulation, machine learning, and data analysis for NASA projects supported by Pleiades
Distributed Active Archive Centers (DAAC)	Process, archive and distribute NASA’s Earth Observing System, a coordinated series of polar-orbiting and low inclination satellites, and related data
Earth Observing System Data and Information System (EOSDIS)	Capture and clean NASA’s Earth science data from remote sensing instruments and curate a multi-petabyte archive distributed across multiple data centers that provide analytics services
Mission data Processing and Control Subsystem (MPCS)	Process Curiosity’s primary data by accessing Mars Reconnaissance Orbiter through NASA’s Deep Space Network, an international array of giant radio antennas
Planetary Data System (PDS)	Archive distributed digital data products gathered from NASA’s planetary missions. Currently, PDS holding data amounts to 700TB, which is available online to the public

Especially, EODIS, one of NASA's major data archive and distribution program, has shown a big leap in its data archive and distribution capability growth over the last few years as shown in Figure 1. With future missions requiring more data storage, EODIS is seeking an innovative solution using open-source cloud native software.

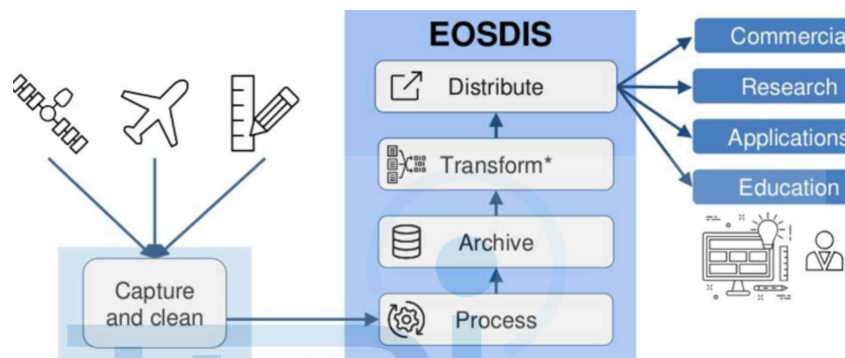


Figure 1. EODIS data system map and evolution¹

The Canadian Astronomy Data Centre (CADC), established in 1986, is a subsidiary of The Canadian Space Agency. Being one of the largest astronomy data centers in the world, CADC houses some of the most important astronomical data collections returned from the Canada-France-Hawaii Telescope, the twin Gemini telescopes and the James Clerk Maxwell Telescope. Every year, CADC delivers over a petabyte data to the global research community using cloud computing and storage technologies and is also at the forefront of astronomical data analysis. Two noteworthy initiatives are summarized in Table 3.

Table 3. CADC's big data projects

Project Title	Objectives
---------------	------------

¹ Data source: <https://www.slideshare.net/AmazonWebServices/stg205earthonaws-how-nasa-is-using-aws>

Common Archive Observation Model (CAOM)	Standardize and unify the core parts of every archive so that a common set of software tools can be used across different archives. Now at version 2.3, CAOM enables the search through metadata returned from the complete set of telescopic data using a single interface.
Canadian Advanced Network for Astronomical Research (CANFAR)	Deliver a platform for very large astronomical datasets enabling the access, process, storage, analysis and distribution. Using its internal architecture which uses cloud-computing and complex algorithms, it supports forefront scientific discoveries.

4.2 Europe (ESCAPE)

The European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures (ESCAPE) is a collaborative cluster project by the European Union which was launched in 2019 under the aim of addressing challenges in open science and data accessibility within pan-European astronomy and particle physics researchers. Leveraging the European Open Science Cloud (EOSC), which is an initiative within the EU to federate existing resources across national data centers, ESCAPE has created a cross-border and multi-disciplinary environment where researchers can be supported with data publication, analytics, advanced computational capacity, and system workflow management. As a result, ESCAPE has developed solutions for big data handling by seven ESFRI facilities and two pan-European research infrastructures. Current ESCAPE outputs can be found in Table 4.

Table 4. ESCAPE Outputs

Project Title	Objectives
Software Repository	Repository of scientific software services used by ESCAPE research facilities
ESFRI Science Analysis Platform	Scientific analysis platform using open access data

Virtual Observatory	High-level astronomical products archive under the aim of discovering using reused data
Data Infrastructure for Open Science	Federated exabyte-scale data infrastructure using an open access data service
Citizen Science	Open gateway for the public through Citizen Science for communicative actions

Besides ESCAPE, individual European country owns its respective astronomical data center. For instance, the UK Astronomical Data Centre provides an open-source data to the public collected from its ground-based telescopes. France's Centre de Données astronomiques de Strasbourg (CDS) also hosts a unique astronomical database and distributes the analytics results to the astronomical community.

4.3 Northeast Asia (NAOJ ADC, KADC)

Finally, Northeast Asia (NEA) also hosts one of the largest astronomical data centers in the world in Japan. Astronomy Data Center (ADC) of the National Astronomical Observatory of Japan (NAOJ) founded in 1995 delivers observational data from its telescopes including Subaru Telescope, ALMA, Nobeyama Radio Telescopes, VERA, ASTE, HINODE, and RISE. Using the collected astronomical data, it conducts various data analysis projects covering a wide range of wavelengths using its own software and high-speed, secure computer network with 40 Gbps speed capacity. NAOJ currently operates three data publication systems listed in Table 5 and two major data analytics projects as shown in Table 6.

Table 5. NAOJ's data publication systems

Project Title	Objectives
NAOJ Data Analysis, Archival and Publication system	A data publication system that enables distribution, analysis and archiving of NAOJ astronomical data and analytics results
JVO system	Virtual observatory portal enabling the public to search reduced astronomical data obtained by NAOJ telescopes

SMOKA	Raw observational data publication system for observatories
-------	---

Table 6. NAOJ ADC's data analytics projects

Project Title	Objectives
Multi-wave-length Data Analysis System	An open-use computing system constructed under the aim of analyzing astronomical data with various wavelengths. Consisted of over hundred different software programs, MDAS helps data reduction as thus enable an effective analysis of astronomical data
Astronomical Data Archives Center	A platform that provides astronomical catalogue services and data archive services to support research activities

Besides NAOJ ADC, the Korean Astronomical Data Center (KADC) under the Korea Astronomy and Space Science Institute (KASI) is also in the infant stage of strategically developing its own unique database which is published in the form of various journals including JKAS, PKAS and JASS. Currently KADC operates a database system for a total of sixteen telescopes, and the collected data is published in different formats for public use. KADC owned observatory data list and publishment status can be found in Table 7.

Table 7. KADC observatory data list and publishment status

Astronomical Data	Data Category	Publishment Status
KMTNet	Visible-light optical spectrum	Fully published
KVN	Radio observation	Fully published
MIRIS	Infrared galaxy imaging	Fully published
SDO	Space environment monitoring	Fully published
VHF Ionospheric Radar	Space environment monitoring	Partially published
e-Callisto	Space environment monitoring	Fully published

5. Analytics tools and organizations

Data Mining, by definition, is the “process of uncovering patterns and other valuable information from large data sets” (IBM, 2021). Due to the following four aspects of astronomical data—volume, variety, velocity and value—effective data mining is becoming the key for astroinformatics to efficiently manage and analyze big data. The vast amount of astronomical data is expected to be generated in the future from multiple Sky Survey Projects as shown in Figure 2. Given this, the role of effective data mining will be more critical.

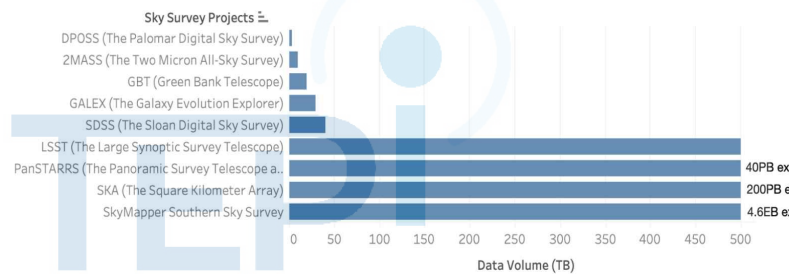


Figure 2. Data Volumes of Different Sky Survey Projects²

Data mining tasks in relation to astronomical data include classification, regression, clustering, outlier detection, and time series analysis, which are the tasks also used in other scientific areas. As a result, an on-going development of analytics tools in the area of astronomy is mostly an application of generic data mining tools to astronomy-specific datasets so that they can be used to overcome some problems faced by the astronomical research society. Three major astronomical data mining softwares are StatCodes, AstroWeka and AstroML.

² Data source: <https://datascience.codata.org/articles/10.5334/dsj-2015-011/>

Table 8. Astronomical Data Mining Tools

Data Mining Tools	Website	Details
StatCodes	http://astrostatistics.psu.edu/stat-codes/	A metasite maintained by the Center for Astrostatistics that provides different statistical codes useful for astronomical data analysis
AstroWeka	http://astroweka.sourceforge.net/	An open-source extension of Weka, a machine learning algorithm toolkit for various data mining tasks, which can analyze data loaded from Astro Runtime, Starlink Tables Interchange Library and Virtual Observatory
AstroML	http://github.com/astroML/astroML	A Python module for machine learning and data mining which aims to support research communities with a module repository for a uniform Python implementations scheme used for statistical data analysis in astronomy

With an increasing importance of analysis of Astronomical data, various international communities are creating astroinformatics and astrostatistics organizations. For instance, the International Statistical Institute (ISI) has founded the International Astrostatistics Association (IAA) in 2012, which is the first global association dedicated only to astrostatistics. In order to promote a keen collaboration between astronomers/astrophysicists and statisticians, and to support astronomers to have better statistical understandings, the IAA holds various conventions and workshops. The IAA now has over 550 members from 56 nations. Even though the majority of astrostatistics and astroinformatics organizations are currently in the form of an interest/working group within its umbrella organization, it is expected that a larger

number of independent global astroinformatics organizations will be formed. All six currently existing organizations are listed in the Astrostatistics and Astroinformatics Portal (ASAIP).

Table 9. Astrostatistics/Astroinformatics Organizations

Organizations	Umbrella Organization	Website
International Astrostatistics Association (IAA)	International Statistical Institute (ISI)	http://iaa.mi.observer.inaf.it/IAA/home.html
Special Interest Group in Astrostatistics	International Statistical Institute (ISI)	https://www.isi-web.org/community/committees/special-interest-groups?id=128
Commission on Astroinformatics and Astrostatistics	International Astronomical Union (IAU)	https://www.iau.org/science/scientific_bodies/commissions/B3/info/
Astrominer Task Force	Institute of Electrical and Electronics Engineers Computational Intelligence Society (IEEE-CIS)	https://asaip.psu.edu/organizations/about-the-ieee-astrominer-task-force/
Working Group in Astroinformatics and Astrostatistics	American Astronomical Society (AAS)	https://aas.org/comms/working-group-astroinformatics-and-astrostatistics-wgaa
Interest Group in Astrostatistics	American Statistical Association (ASA)	http://astrostat.org

6. Challenges in Astroinformatics

6.1 The Education Gap

Even though the fields of Astronomy and Astrophysics have strong needs for data analytics skills, the biggest challenge that has been faced by the community is the education gap. Astronomers and astrophysicists who have been utilizing traditional research methodologies relying heavily on observational findings lack understanding of big data preparation, advanced statistical analysis, and computational algorithms, thus prohibiting them from fully utilizing these useful technologies. This has also been emphasized by Feigelson's 2016 paper, where he mentioned, "In many papers, the analysis is confined to a narrow suite of familiar statistical methods. Even these traditional methods can be misused" (Feigelson, 2016). This educational gap is primarily due to weak statistical training in the graduate and undergraduate education of astronomers where courses in advanced statistics, applied mathematics and computer science are not a part of the required curriculum. Consequently, professional statistical training should be provided for non-student and professional astronomers beyond informal on-the-job training. Additionally, college curriculum should incorporate data science courses for students pursuing astronomy and astrophysics.

6.2 The funding gap

Funds granted to institutions and researchers in the field of astroinformatics and astrostatistics are very scarce. For instance, in the U.S., the "NSF has had short-lived interdisciplinary grant programs to promote mathematical developments for astronomy. ... While astronomers do have access to agency-wide programs in cyberscience such as the Computational and Data-Enabled Science and Engineering (CDS&E) and various cyber infrastructure programs, the success rate of such proposals may be de facto limited by level of buy-in from [the] NSF's Division of Astronomical Science" (Eadie, 2019). Consequently, large-scale programs that help to build astronomical data analytic infrastructure such as data and computational centers need to be funded at a long-term, stable, and scalable level.

7. Conclusions and strategic recommendations

In conclusion, this paper has deep dived into the status quo of big data usage in the modern fields of astronomy and astrophysics by investigating global space agencies' astroinformatics initiatives, data analytics tools and organizations, and the challenges faced by the research society. Big data is becoming a major method of astronomical investigation, as astronomical data collection technology is growing at a very rapid pace. As a result, more and more research initiatives are going beyond the traditional boundaries of simple statistical analysis and are implementing high-level analytical methods such as computational algorithms and machine learning techniques. However, due to constraints such as educational gap and lack of proper funding to support astronomers and astrophysicists, the rapidity of growth in data analytics is relatively slow and has led to an imbalance between data collection and data analysis.

Accordingly, a strategic plan that supports the development of data analytics tools and technology is necessary, so that massive data collected from advanced facilities can be translated into knowledge in an efficient yet effective manner. Zhang and Zhao also emphasized the importance of this in their 2015 paper by saying, "each step of data to knowledge from data generation, data collection, data transformation, data storage, data management, data preprocessing, data mining, data visualization, data understanding, data evaluation, and data explanation depends on the improvement of existing tools and technologies or discovering new tools and technologies" (Zhang & Zhao, 2015). Additionally, a platform that enables collaboration among computer scientists, data statisticians, astronomers and astrophysicists is crucial for fostering interdisciplinary study opportunities, which has the potential to generate synergetic effects resulting in unique findings.

References

CADC. 2019. Canadian Astronomy Data Centre research facility. *National Research Council Canada*. Retrieved at:

- [<https://nrc.canada.ca/en/research-development/nrc-facilities/canadian-astronomy-data-centre-research-facility>]
- Eadie, Gwendolyn, *et al.*, 2019. Realizing the potential of astrostatistics and astroinformatics. *Astro2020 State of the Profession Consideration White Paper*.
- ESCAPE. Project ESPACE – About Us. *ESCAPE*. Retrieved at: [<https://projectescape.eu/about-us>]
- Feigelson, Eric. 2016. The changing landscape of astrostatistics and astroinformatics. *Astroinformatics Proceedings IAU Symposium*, No. 325. DOI: 10/1017/S1743921317003453
- IBM. 2021. Data Mining. *IBM Cloud Education*. [<https://www.ibm.com/cloud/learn/data-mining>]
- Meyer, Eileen. 2018. Big Data is Transforming How Astronomers Make Discoveries. *Smithsonian Institution*, Retrieved at: [<https://www.smithsonianmag.com/science-nature/next-big-discovery-astronomy-scientists-probably-found-it-years-ago-they-dont-know-it-yet-180969073>]
- NAOJ. Astronomy Data Center – About Astronomy Data Center. *NAOJ*. Retrieved at: [<https://www.nao.ac.jp/en/research/project/adc.html>]
- NASA. NASA Quantum Artificial Intelligence Laboratory (QuAIL), *National Aeronautics and Space Administration*, Retrieved at: [<https://ti.arc.nasa.gov/tech/dash/groups/quail/>]
- SDSS. The Sloan Digital Sky Survey: Mapping the Universe. *SDSS*, Retrieved at: [<https://www.sdss.org>]
- Shannon, Michael. 2019 How does NASA use big data? *Bigdata-madesimple*. Retrieved at: [<https://bigdata-madesimple.com/how-does-nasa-use-big-data/>]
- SKAO. The SKA Project. *SKAO*, Retrieved at: [<https://www.skatelescope.org>]
- Wynn, Renee, *et al.* (eds.). 2018. *The Big Data Wave*. NASA IT Talk
- Zhang, Yanxia and Zhao, Yongheng. 2015. Astronomy in the Big Data Era. *Data Science Journal*, 14. P.11. DOI: [<http://doi.org/10.5334/dsj-2015-011>]

15 / Astronomy in the Era of Big Data: The Status Quo, Challenges, ...

Min Jun Kim

18477경기도 화성시 동탄대로시범길 20 #1413-1201

E-mail: minjunkim129@gmail.com

접수일자: 2021. 07. 27.

게재일자: 2022. 07. 17.

