An Acad Bras Cienc (2021) 93(Suppl. 1): e20200861 DOI 10.1590/0001-3765202020200861
Anais da Academia Brasileira de Ciências | *Annals of the Brazilian Academy of Sciences*
Printed ISSN 0001-3765 I Online ISSN 1678-2690
www.scielo.br/aabc | www.fb.com/aabcjournal

MATHEMATICAL SCIENCES

Data Science Strategies for Multimessenger Astronomy

REINALDO R. ROSA

Abstract: This article aims to identify and suggest data science strategies to strengthen scientific research in astronomy. The improvements in data workflow performance that can be provided by these strategies can be crucial to the multimessenger astronomy (MMA). A special focus is given to the treatment of raw data in the context of *big data networks* for BRICS astronomy initiatives. A preliminary design of a prototype that incorporates an MMA data cube into a data lake system is presented.

Key words: Astronomical data science, big data, data lake, data cube, MMA.

INTRODUCTION

One of the biggest challenges for researchers today is accessing and analyzing large volumes of data known as *big data*[1]. Such huge data-sets are generally obtained through the *internet of things* (*IoT*) (Ahmed et al. 2017) and, therefore, stored mainly in the digital cloud. In most cases, information technology allows receiving and storing large volumes of heterogeneous data in *digital warehouses*, databases that can be consulted via the Internet. The concept of data warehouse allows scientific data to be made available in an organized way to facilitate its access to scientific research activities. From a scientific data warehouse, it is possible to quickly identify a set of data of interest, download[2] it to provide data mining in search of information that can generate redundant or new knowledge. Therefore, for each scenario of scientific research based on data mining, the main strategy is to understand how the data life cycle works and how we can put data science models into production more efficiently, so that they begin to replicate usefully and effectively among users. For this, it is necessary to identify each of the main stages of the data work-flow: production, storage, access and acquisition, use and disposal. In this context, one of the most important strategies for 21st century astronomy, which are virtual observatories (VO) (Carvalho et al. 2010, Allen et al. 2019), should incorporate data life cycle management in an accurate and efficient manner.

In practice, the data life cycle can be modeled in different ways. The basic form includes a virtual cyclical flow of uses and reuses, which allows for reformulations in its content, structure and format (Kelleher & Tierney 2018). Feedback strategies are best suited for genuine observational data that can retain value almost indefinitely. In astronomy research, for example, the data generated by a given generation of telescopes (especially for large surveys) can become important for

[1]See the Appendix for the 4V concept of Big Data in Astronomy.
[2]This also includes the action of just relocating in the cloud.

calibrating the next generation. Therefore, the value and reuse of data must be improved through investments in information technology infrastructures, with an emphasis on digital management and maintenance. Thus, the data life cycle requires the data management system to determine which data should be kept at some level, specifying the reasons and for how long. In the scope of multi-messenger astronomy, it is essential that the concatenation of data and the triggering of alerts are also properties of the cycle. As the cycle period can be crucial for the objectives of triggering complementary observations to be redirected, analytic access to the raw data may be necessary. In this scenario, the *data lake* concept gains importance and can be a fundamental strategy (Li et al. 2017).

This manuscript is organized as follows. Next section presents a list of strategies, within the scope of data science, which aims to be effective in the context of astronomy. Also in this section, multi-messenger astronomy is highlighted and the main strategies for it, as *data lake*, are identified. Third section presents a *data lake* solution for multi-messenger astronomy in the context of the BRICS. The last section presents the main concluding remarks of this study. The concept of big data astronomy is discussed in the Appendix.

DATA SCIENCE STRATEGIES FOR ASTRONOMY

Today we can say that astronomy, at its most professional end, is a science whose most prominent discoveries are driven by a large amount of data (observed through telescopes or simulated from HPC) that fits within the contemporary concept of Big Data (see Appendix).

When we take into account the data production rate of large telescopes[3] we realize that massive raw data will be collected almost continuously. In this scenario, a new list of data science strategies emerges that allows to streamline the scientific production in astronomy. In this framework, big data has no chance of being useful if astronomers try to process it using traditional mechanisms, such as data warehouses and conventional data analysis techniques (Li et al. 2017). Therefore, the main strategies include technological solutions for generating, storing, extracting and mining information in an agile and effective way. All of them include software solutions (machine learning algorithms, frameworks and virtual platforms), hardware (intelligent sensors, supercomputing based on networked minicomputers, HPC based on large data stores) and hybrid solutions such as embedded software using FPGA (Jin & Finkel 2018). In such a broad scenario, the so-called *data lake* concept is an information system that has three main characteristics: (i) an enormous scalable data storage capacity; (ii) a self-contained data mining center and (iii)automatic security modules for ETL[4].

Considering the major projects of observational astronomy that are coming on or will come on stream in the next decade (2021-2031) (see Appendix A), we can identify and list the main demands involving strategies directly related to data science:

1. Development of a *Data Cube* prototype (Cuzzocrea 2010) for managing the life cycle of data adapted for both physical

[3]for example, from the Square Kilometer Array (SKA) approximately 150 GB of data per hour will be transmitted from each satellite dish to a data processor hub.

[4]System for *Extract*, *Transform* and *Load* data from one source environment to another where data must be ready to be used in applications.

and virtual observatories where meta-data cleaning solutions and dissemination of high quality data are mandatory. The use of NoSQL solutions (as Haddop and Spark) is crucial (Buchschacher et al. 2019);

2. Integrating globally distributed resources (as Astronomical Database) into a unified cyber platform[5] as, for example, fast algorithms for astronomical data reduction, data mining and visualization (Giommi et al. 2020);

3. Design of distributed cloud based engines for knowledge discovery in massive archives of astronomical data cubes (Teuben et al. 2019);

4. Development of Data Science Center based on Open Scientific Resources where the aggregation and linking of observational metadata in the ADS is of extremely importance (Teuben et al. 2019);

5. Development of Agile (Fast, Cheap and in the Cloud) Astronomical Data Archives (Teuben et al. 2019);

6. Design and application of astronomy Data Lake where smart data pipelines and real-time processing are available (Barchi et al. 2020);

7. Development of a data science framework aimed at the dissemination, outreach and promotion of citizen science for the efficient treatment of big data in astronomy. In addition to working in search and classification missions, the production of increasingly comprehensive and complete reference catalogs depends on the effectiveness of this strategy (Marshall et al. 2015);

All of these strategies must consider, always within the data life cycle paradigm, knowledge outputs that should be interpreted as long-term targets. Some of them are as follows: Affordable immersive visualization of astronomical data and innovative user interfaces; Improvements of astrophysics source code libraries especially in Python, R and their associations with parallel programming (eg pyCUDA for GPUs); future astronomical data formats from data cubes innovation including monitor web-analysis systems (Jin & Finkel 2018, Teuben et al. 2019, Barres de Almeida 2020, Zhang and Zhao 2015).

The list of strategies presented above is not intended to be complete and the author understands that it represents an initial kick to stimulate its discussion in order to seek the contribution of the community to build it in an increasingly realistic workfkow. Its content comes mainly from two sources. First, based on the guidelines outlined in the last five years by the main astronomy associations (with emphasis on ASP, AAS and IAU) and, second, by the experience acquired by the author when participating, as coordinator and member since 2017, of the SAB's Information Technology Commission[6](Barres de Almeida 2020).

THE MULTI-MESSENGER ASTRONOMY CONTEXT

In the mid-20th century, new observatories have been able to add new windows for observing the universe to the list of effective astronomical sources. These messengers in addition to the electromagnetic waves now include particles (for example, neutrinos and cosmic rays) and space-time fluctuations interpreted as gravitational waves. Thus, a new frontier in astronomy has been forming

[5]These facilities are not restricted to the concept of VO, but also incorporate more general portals, hubs, colabs and repositories.

[6]Brazilian Astronomical Society (SAB).

from the concatenated study of several wavelengths combining observations from different bands of the electromagnetic spectrum. After the 2017 canonical event[7] the name Multi-messenger Astronomy (MMA) was coined by the astronomical community because of the use of the new types of cosmic *messengers* to study the physical processes of deep space (Bartos et al. 2017). In this new scenario, the great challenge that drives MMA, as a new area of astronomy, is to effectively combine information from each type of source to provide a unified understanding of the phenomenon. Therefore, in addition to new instrumental and logistical aspects, astronomers are beginning to discuss what are the main challenges, strategies and opportunities in data analysis for multiple messenger astrophysics.

Data Science for MMA

MMA is on the frontier of astronomy and therefore, from the point of view of data science, it is subject to all the strategies presented in the previous section. However, it has a property that stands out compared to conventional astronomy. It is about the need for agile concatenation between the observations of different messengers, which still come from different telescopes distributed in different locations. This demand includes the special needs listed below, ranging from instrumental concatenation (preferably autonomously) to the production (also autonomous) of knowledge: (i) Autonomous telescope operations and scheduling; (ii) Local and global cloud infrastructure for processing and storage MMA data; (iii) Delivering accessible and science-ready MMA data; (iv) Data discovery across heterogeneous datasets based on machine and deep learning; and (v) promoting meetings with data science experts to solve problems, improve techniques and advance the data life cycle[8]. We conjecture here that the most suitable format for MMA data within the main needs (following a workflow) presented above lies in the data cube paradigm.

Thefore, the proposed MMA data cube (MMA-DC) follows the multidimensional paradigm of online analytical processing (OLAP) defined for the 3 first order dimensions (time (z), neutrino data (x) and gravitational wave data (y)) allowing to identify in each element a second order data cube (images, time series or spectra) for sub-intervals of time along z. In this way, second order data cubes are also identified in each element of the first order cube. That is, both first and second order cubes are indexed in 3 dimensions. Usually, MMA-DC OLAP are executed more quickly than other approaches, mainly because it is possible to index directly in the data cube structure to collect subsets of data.

Based on the immersion of a specific MMA data cube within the Data Set Resources layer of a data lake, a unified conception for the needs pointed out in this section is presented in the next one.

A Data Lake Resource for BRICS MMA

Collaboration is ongoing between the BRICS countries to create a transient and multi-messenger network bringing together existing and future observation infrastructures on the MMA frontier. Within this scope, it is opportune to propose here a first step towards a

[7] An irrefutable message of gravitational waves was finally observed in 2017. It has been proven that this signal occurred due to the fusion of two neutron stars about 130 million light years from here. Almost simultaneously – just two seconds after the gravitational wave signal arrives – the first electromagnetic signal (in the form of gamma rays) arrived, representing the detection of the first robust multi-messenger signal involving gravitational waves.

[8] The ideal in this activity would be to incorporate cooperative-competitive scientific marathons known in the business field as *hackathons*. Perhaps more appropriate in the academic segment is the use of *sci-hachathon*.

innovative data processing pipeline based on a MMA data cube embeded in a data lake system, that can be discussed by the BRICS astronomy community.

Data lake (DL) is an active and flexible data system, for large quantities and varieties of data, both structured and unstructured, providing a schema-less repository for raw data with a common access interface. The concept proposes greater independence from standardization and greater flexibility in modeling, resulting in almost unlimited potential for operational diversity in data mining (Ivezić et al. 2020). Thus, as the volume of data, the variety of data and the richness of metadata increase, the dependency on a single workflow decreases. This inverse relationship takes the rigidity out of ETL-warehouse paradigm and makes data science agile. Therefore, when we think of observational concatenation for the agile analysis of a very wide and diverse dataset, we need to take into account the access and mining of raw data. In this context, it is desirable that the concept of data lake be incorporated into the multi-messenger astronomy scenario. We therefore started from the DL prototype presented for a VO platform in China (Li et al. 2017). Our prototype simplifies the administration and security module, and incorporates a data cube in the storage layer. The embedded data cube proposes to semi-structure raw data within the MMA paradigm (see Figure 1).

The system architecture is shown in Figure 1. The administration and security module supports two agile software solutions: *Cloudera* or *Yarn*. The storage base incorporates all the benefits of the *GlusterFS* platform (unlimited storage, backup and automatic recovery nodes). The analytical module is based on google COLAB with Keras and Tensor Flow for Machine and Deep Learning applications already used by the astronomical community (Rosa et al. 2018, Barchi et al. 2020, Ivezic et al. 2020). The data and information flow pipelines are made up of data quality resources, simplified NoSQL ETL and a Unformatted System Services (USS) (ex. *z*/*OS*) for documentation. This economy of layers presents an initial prototype that will be tested as a practical DL solution that can be used as the basis for several astronomical DL projects, as it provides a flexible and extensible structure for continuous data entry in different formats (.txt,. csv, .arff, .tiff, .jpg, .jpg, .fits). Naturally, the data lake feature presented is not restricted to MMA. It can be adapted, generalizing the MMA data cube, for astronomy as a whole.

At the current stage, all modules have been independently tested by a development team. We are starting the integration phase, which, after being completed, will be tested in two radically different environments: (i) GoogleColab and (ii) as software embedded, for example, in a Raspberry Pi4 minicomputer. In the future, the Project prototype provisionally called DLMMAbr, will be mirrored in the project of the same name already created on GitHub. Its integration with platforms for astronomy (such as AstroCloud and OpenUniverse) will also be considered strategic. Cooperative initiatives within the BRICS will be encouraged.

CONCLUSIONS

In addition to the MMA data lake concept presented in this work, the general strategy to be developed asap is to materialize software and hardware solutions that allow to combine, in an agile and concatenated way, observations and data from the entire electromagnetic spectrum with gravitational waves and neutrinos. Because these data come from heterogeneous detector networks and configurations, their analysis is often critical in terms of temporal triggers to guide and concatenate data analysis from

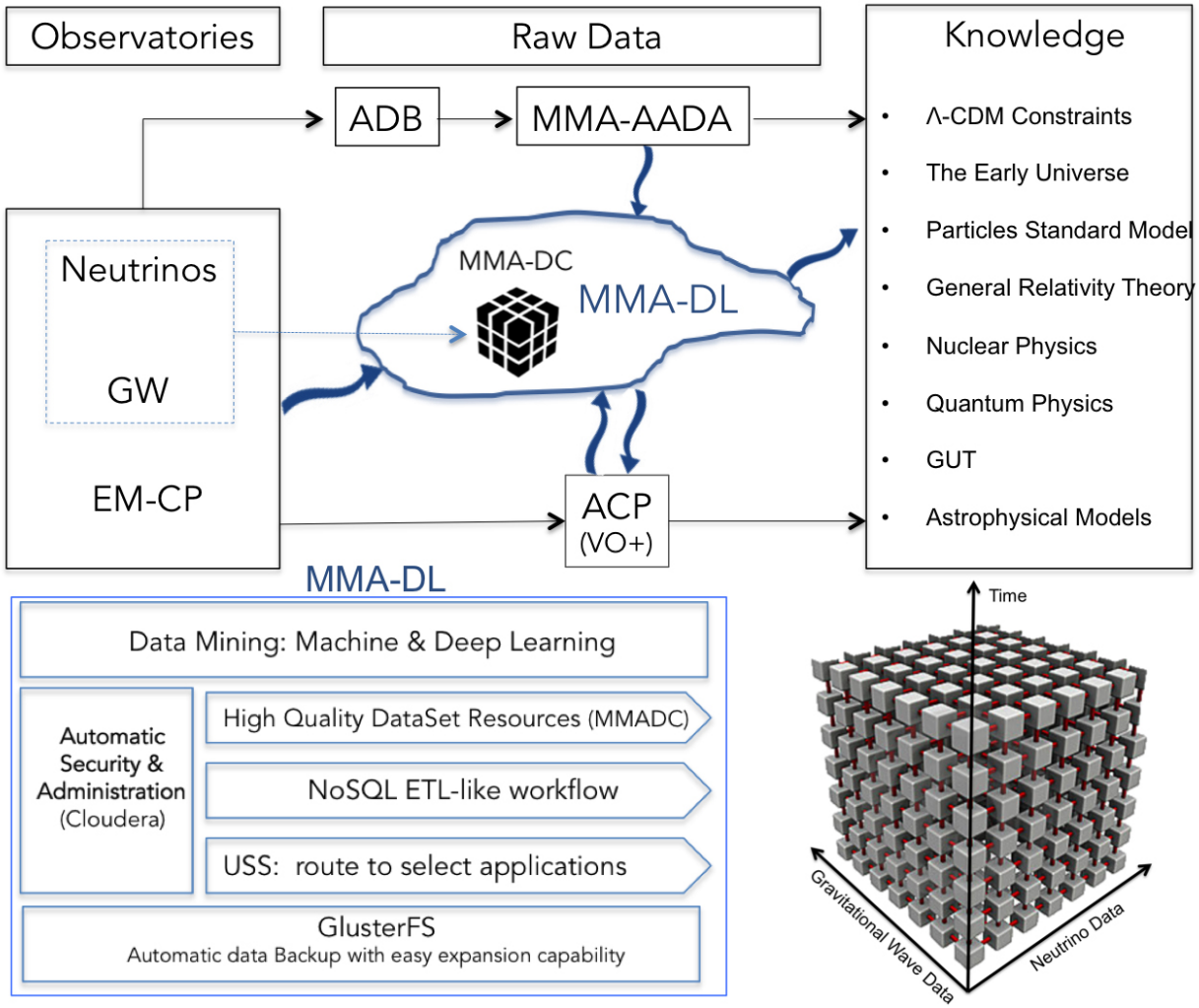

Figure 1. Design of a data lake prototype for MMA with minimally structured raw data within the OLAP paradigm for multidimensional data cubes. The acronyms represent the following abbreviations: ADB for Astronomical Data Base; ADA for Agile Astronomical Data Archive; ACP for Astronomical Cyber Platforms; EM for Electromagnetic Counterparts and GW for Gravitational Waves. *Raw data in this figure must be understood within the ETL paradigm reporting all dataset not yet incorporated into a data warehouse-like repository.*

multiple observations. Therefore, it is in practice a major challenge that the related sciences and technologies will face when providing the full use of data from MMA.

Roughly speaking, we are referring to a new area in which scientific capabilities depend on the interaction between observation, theory and computational work with a strong dependence on instrumental and computational modeling. In this context, the advances in data science assisted by the advances in high performance computing (with emphasis not only on software and hardware, but also on peopleware) present a new universe of opportunities for the entire scientific community related to the BRICS astronomy iniative.

Finally, from a practical point of view, it is important to reinforce peopleware strategies. In this matter, two new initiatives of practical activities of data science were highlighted here: sci-hackathons and citizen science. The latter already in practice in astronomy where we

take as examples the success of the *Galaxy Zoo*, *EisnteinHome* and *Planet Hunters TESS* initiatives.

Acknowledgments

The author thanks the development team constituted until now by L.A. Filho, P.H. Barchi, R. A. Sautter, R.A. Sych, M. Bento, M.A.U. Cintra, W. Chun, N. Joshi and P.Z. Giovanni. Part of the analytical tools presented in this article was partially funded by the São Paulo State Research Support Foundation (FAPESP, project number 2014/11156-4).

REFERENCES

AHMED E, YAQOOB I, HASHEM IAT, I KHAN I, AHMED AIA, IMRAN M & VASILAKOS AV. 2017. The role of big data analytics in Internet of Things. Comput Netw 129(2): 459-471.

ALLEN MG, DOWLER P, EVANS JD, CUI C & JENNESS T. 2019. The International Virtual Observatory Alliance, proceedings of Astronomical Data Analysis Software and Systems XXVIII published by ASP: arxiv.org/abs/1903.06636.

BARCHI PH, DE CARVALHO RR, ROSA RR, SAUTTER RA & SOARES-SANTOS M. 2020. Machine and Deep Learning applied to galaxy morphology-A comparative study, Astronomy and Computing 30: 100334.

BARRES DE ALMEIDA U, KRONE-MARTINS A, DIAZ MP, DO NASCIMENTO JD, LEO WV, ROSA RR, SAITO RK. 2020. Information technology & astronomical data in Brazil: Perspectives and proposals. Boletim da Sociedade Astronômica Brasileira 32(1): 142-146.

BARTOS I, KOWALSKI M & MULTIMESSENGER ASTRONOMY. 2017. IOP Publishing, Bristol, UK, ISBN: 978-0-7503-1369-8.

BUCHSCHACHER N, ALESINA F & BURNIER J. 2019. No-SQL Databases: An Efficient Way to Store and Query Heterogeneous Astronomical Data in DACE 523, Astronomical Data Analysis Software and Systems XXVIII, 405, ASP.

CARVALHO RR ET AL. 2010. The Brazilian Virtual Observatory - A New Paradigm for Astronomy. J Comp Int Sci 1(3): 187-206.

CUZZOCREA A. 2010. OLAP Data Cube Compression Techniques: A Ten-Year-Long History, Lecture Notes in Computer Science, Springer.

GIOMMI P, ARRIGO G, BARRES DE ALMEIDA U, DE ANGELIS M, DEL RIO JV, DI CIACCIO S, DI PIPPO S, IACOVONI S & POLLOCK A. 2020. The Open Universe Initiative, in Space Capacity Building in the XXI Century, Springer, p. 377-386.

IVEZIĆ Z, CONNOLLY AJ, VANDERPLAS JT & GRAY A. 2020. Statistics, Data Mining, and Machine Learning in Astronomy A Practical Python Guide for the Analysis of Survey Data: https://press.princeton.edu/books/hardcover/9780691198309/statistics-data-mining-and-machine-learning-in-astronomy.

JIN Z & FINKEL H. 2018. Power and performance tradeoff of a floating-point intensive kernel on OpenCL FPGA platform, IEEE Symp on Parallel and Distributed Processing, p. 716-720. doi:10.1109/IPDPSW.2018.00115.

KELLEHER JD & TIERNEY B. 2018. Data Science, MIT Press Essential Knowledge series.

KUHN TS. 1996. The structure of scientific revolutions, Univ. Chicago Press.

LI C ET AL. 2017. The Design and Application of Astronomy Data Lake in China-VO, Astronomical Data Analysis Software and Systems XXV ASP Conference Series, Vol. 512, In: Nuria P,Lorente F, Shortridge K & Wayth R (Eds), Astronomical Society of the Pacific.

MARSHALL PJ, LINTOTT CJ & FLETCHER LN. 2015. Ideas for Citizen Science in Astronomy, Ann Rev Astron Astrophy 53: 247-278.

ROSA RR, DE CARVALHO RR, SAUTTER RA, BARCHI PH, STALDER DH, MOURA TC, REMBOLD SB, MORELL DRF & FERREIRA NC. 2018. Gradient pattern analysis applied to galaxy morphology. Mon Notices Royal Astron Soc Lett 477(1): L101-L105.

TANSLEY S & TOLLE KM. 2009. The Fourth Paradigm: Data-intensive Scientific Discovery, Microsoft Research.

TEUBEN PJ, POUND MW, THOMAS BA & WARNER EM. 2019. Astronomical Data Analysis Software and Systems XXVIII, V. 523, eISBN: 978-1-58381-934-0.

ZHANG Z & ZHAO W. 2015. Astronomy in the Big Data Era. Data Sci J: doi:10.5334/dsj-2015-011.

How to cite
ROSA RR. 2021. Data Science Strategies for Multimessenger Astronomy. An Acad Bras Cienc 93: e20200861. DOI 10.1590/0001-3765202020200861.

REINALDO R. ROSA
https://orcid.org/0000-0002-2962-4322

Lab for Computing and Applied Mathematics (LABAC), National Institute for Space Research (INPE), Av. dos Astronautas 1758, 12245-690 São José dos Campos, SP, Brazil

E-mail: rrrosa.inpe@gmail.com

Manuscript received on June 3, 2020; accepted for publication on June 24, 2020

APPENDIX: BIG DATA IN ASTRONOMY

The concept of epistemological paradigm indicates how the advancement of scientific knowledge has evolved over the history of civilization (Kuhn 1996, Tansley & Toole 2009). According to some important contemporary scientists, science in the 21st century is experiencing the shift to a fourth paradigm after a sequence of three: (i) science guided mostly by empirical experiments, (ii) science guided mostly by theory and (iii) science guided mostly by computer simulation. The fourth paradigm suggests that new and more important scientific discoveries are (and will be more and more) strongly conditioned by the intensive use of large volumes of data called *Big Data* (Kelleher & Tierney 2018). The science driven by big data is the 4th paradigm, interpreted today as *Data Science* by most of the productive and academic sectors.

Many authors try to define *Big Data* from the so-called set of the four basic properties (4V): *volume*, *velocity*, *variety* and *value*. Properties that are related to the production, transmission, storage, acquisition and analytical manipulation of digital information. In practice, the workfow velocity (in bytes per unit of time) is a very important quantitative factor. In this way, *big data* are large data sets that are generally heterogeneous in their variety and value and, therefore, demand great storage and processing power throughout their workflow.

Nowadays, the main source of *Big Data Workflows* is the active network of electronic devices connected to the internet, named by Kevin Ashton, in 1999, by the name of the *Internet of Things* or simply *IoT* (See Figure 2). In the context of *IoT*, the search for information or new knowledge, whether working with simulations, measurement sensors or already published data, is based on large amounts of data that are, for the human brain, very difficult to store, manipulate, analyze and understand. It is in this context that *data mining* (mostly based on techniques from *machine learning*) is one of the fundamental gears in the process of extracting information from *big data*.

Astronomical observatories generate an impressive amount of data. For example, ALMA (Atacama Large Millimeter Array), operating in Chile, adds about 2 TB of data to its files every day. And as we know, each generation of observatories is usually at least ten times more sensitive than the previous one, either because

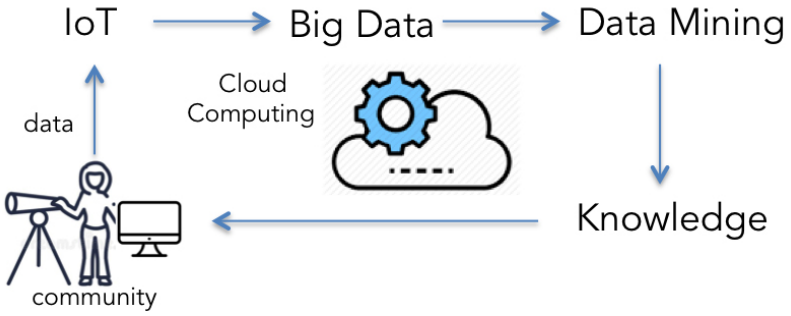

Figure 2. Today, the data life cycle can follow the path defined by data science based entirely on cloud computing. The observer makes use of IoT devices to place their data in the cloud. The IoT is generating Big Data where your data is now inserted. A user can mine the data and extract knowledge directly in the cloud. This knowledge can serve to generate new data and thus boost the cycle.

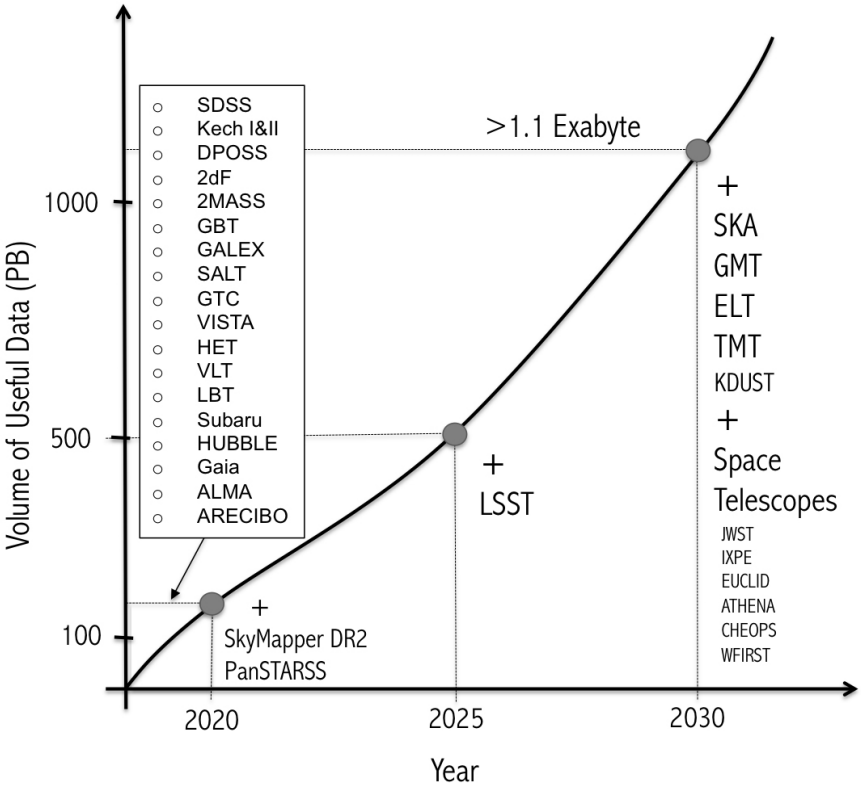

Figure 3. The graph shows the evolution of useful data generated from major astronomical observational projects. Useful data, in the context of data science, are those to be extracted automatically from the raw data workflow which, in technical language, is called Data Lake. These values are generally inferred from the estimates provided for each instrument under the condition of full operation. In the new Data Lake approach, it is reasonable to consider that up to 25% of raw data will be discarded for several reasons involving pre-processing parallelization, quality or priority.

of improved technology or because the mission is simply greater. Depending on the duration of a new mission, it can detect hundreds of times more astronomical sources than previous missions on the same wavelength. With LSST operating from 2023, adding to the continuous production of data from the observations in progress, this load could exceed in 2025 the 500 petabyte mark. By 2030, in addition to SKA, other projects of gigantic size, where some have the direct participation of the BRICS countries will also be concluded and activate. After that date, considering a significant number of new space telescopes already in operation, the equivalent of a data deluge is about to occur. In fact, as shown in Figure 3, by 2030 Astronomy will surpass the mark of 1.1 exabytes of useful data.

Therefore, new practices in data science and information technology are crucial to address the big data problems facing 21st century Astronomy. As in other fields of science, in the era of *Big Data* and the IoT, Astronomy has shown significant innovations with emphasis on solutions in astroestatistics and astroinformatics, in addition to innovative resources such as the VO network (Allen et al. 2019). In this way, the concept of *data science* in Astronomy is consolidated, as well as its relationship with information technology mainly that based on the cloud computing resources.