CHAPTER 2

ASTRONOMICAL DATA

Hulusi GÜLSEÇEN*, Hasan H. ESENOĞLU**

*Asos. Prof., İstanbul University, Science Faculty, Astronomy and Spaces Sciences Department, İstanbul, Turkey
E-mail: hgulsecen@istanbul.edu.tr

**Asos. Prof., İstanbul University, Science Faculty, Astronomy and Space Sciences Department, İstanbul, Turkey
E-mail: esenoglu@istanbul.edu.tr

DOI: 10.26650/B/ET06.2020.011.02

Abstract

Space telescopes have increased the quality of data collection for today's astronomy. In parallel to this, obtaining high quality data with high technology and good resolution focal plane detectors in accordance with the developments in material science in the ground-based observations has been achieved. With the new generation of ground based and space observations, global campaigns also brought continuity in data acquisition and increased performance. Finally, the fact that theoretical outputs can be made to allow in today's technology, for example, the detection of gravitational waves in the universe and these add new ones to the existing data. In addition, there has been a significant increase in data archiving, reduction and processing together with the number and variety of data collection tools. Astronomers have been able to overcome the facilitation in these processes in their own way: manpower waste has been reduced with autonomous telescopes, the data has been transformed into informatics (astroinformatics) with pipelines, the workload has been reduced to large masses by establishing a virtual observatory, and finally smart applications have been opened with the provided big data and new open areas have been reached with a future such as data mining. In this way, there has been progress in solving many astronomical events in the universe. This chapter is orginized in two subsections. In first, we are discussing how to solve problems in astronomy by using big data. In the second, we mention about big data sources in astronomy. The importance of data in astronomy, sources of data, big data in regards to the discovery of universe and analyzing data are the topics discussed in these subsections.

Keywords: Astroinformatics, Astrostatistic, Astronomy, Big data, Machine learning, Processing, Reduction

1. Introduction

Astronomy is the study of physics, chemistry, and evolution of celestial objects and phenomena that originate outside the Earth's atmosphere, including supernova explosions, gamma ray bursts, and cosmic microwave background radiation (Zhang and Zhao, 2015). Since astronomy is a science that studies celestial bodies, the objects in space can only be investigated by examining the light coming or reflected from them. Thus, the only source astronomers have is light.

There are many difficulties when investigating a celestial body. The Earth, the Sun and the Solar system are constantly in motion. Also, more distant celestial bodies such as stars and galaxies are constantly in motion. For this reason, the location of a celestial body at the time of observation, the position of the earth and the time of observation are very important.

Studies with celestial bodies must be reduced to a heliocentric coordinate system. Observation time (which is one of the main parameters of astronomical data sets) should also be reduced to HJD (Heliocentric Julian Day).

The time and the coordinates of both the celestial body and the detector (telescope, satellite, CCD, etc.) are indispensable parameters of a data set regardless of the wavelength in which the field of astronomy is studied.

We can roughly divide astronomy into three areas of study. These are astrometric, photometric and spectroscopic studies. Roughly, we can classify the celestial objects to be observed as the sun, the objects of solar system, stars, Milky Way, galaxies, and galaxy groups. These celestial bodies are observed with different devices at different wavelengths of the electromagnetic spectrum. The classes of astronomy in terms of wavelength can be made as follows: gamma-rays astronomy, x-rays astronomy, ultraviolet astronomy, optical astronomy, infrared astronomy and radio astronomy.

Astrophysics is the branch of astronomy that studies the physics of the universe, in particular, the nature of celestial objects rather than their positions or motions in space. Astrophysics typically uses many disciplines from physics, including mechanics, electromagnetism, statical mechanics, thermodynamics, quantum mechanics, relativity, nuclear, particle physics, and atomotic and molecular physics to solve astronomical issues (Zhang and Zhao, 2015).

The occurrence times and life span of the events taking place in space also vary greatly. For example, gamma-ray bursts last for a few seconds while solar eruptions and binary star

eclipses last for a few minutes and several hours to years, respectively. The lives of stars last from ten million years to several billion years.

Gamma-ray bursts (GRBs) in Astronomy are flashes of gamma-rays associated with extremely energetic explosions that have been observed in distant galaxies. They are the brightest electromagnetic events known to occur in the universe after the big bang. Bursts can last from milliseconds to several minutes. The initial burst is usually followed by a longer-lived afterglow emitted at longer wavelengths (x-ray, ultraviolet, optical, infrared, microwave and radio). Targets of Opportunity (ToO) are astronomical objects undergoing unexpected/unpredictable transient phenomena and proposed for observation. The observations are normally urgent because of the transient nature of the event and may require even an immediate intervention at the telescope. ToO include objects that can be identified before the onset of such phenomena (e.g. dwarf novae, x-ray binaries) as well as objects which cannot be identified in advance (e.g. novae, supernovae, gamma-ray bursts). Modules have been developed for fast telescopes that respond to GRB alerts robotically in collaboration with the coordination of data networks. An example of deployed T60 at the TUBITAK National Observatory (Antalya, Turkey) was carried out by embedded software of the robotic telescope (Dindar et al., 2015). The telescope responds to GRB triggers transmitted from the Goddard Space Flight Center alert system thanks to this autonomy. It uses the Gamma-Ray Explosion Coordinates Network - GCN (formerly known as the BATSE Coordinates Distribution Network, BACODINE) while doing this. There are also some pipelines designed for Gaia alerts (http://gsaweb.ast.cam.ac.uk/alerts/alertsindex) similar to GRB alerts. One of these is “AlertPipe” which is responsible for real-time detection and classification of anomalies and transient astrophysical phenomena. The pipeline works within the Gaia data processing stream.

Recent advances in satellite and CCD technology have allowed for a more detailed examination. Dark energy, dark matter and exoplanet research have been accelerated thanks to these developments in technology.

Advances in computer technology, the enormous expansion of new storage capacities, the diversity and organization of astronomical data have led to the addition of two new fields of study to astronomy. In particular, data mining, machine learning and artificial intelligence applications have started to be used in astronomy studies.

Finding solutions to the problems in astronomy with big data and subjects of big data in astronomy are discussed under the relevant subheadings below.

2. Solving Problems in Astronomy with Big Data

Statistics plays an essential role in data-rich astronomy. Scientific insights cannot be extracted from massive datasets without statistical analysis. The statistical challenges are not simple; image analysis, time series analysis, nonlinear regression, survival analysis, and multivariate classification are all critically important (Feigelson and Babu, 2012).

Data in a method called DFS (Distributed File System) is placed wherever there is a free computer on Earth. For example, a part of the picture you upload to Facebook can be held on a computer in China and the other part can be held on a computer in Canada. Hadoop combines these two pieces of information in milliseconds when you click to view.

Astronomy was developed in two main areas namely "Astrostatistics" and "Astroinformatics". Astrostatistics can be summarized as the application of the science of statistics to the sciences of astronomy and astrophysics. Astroinformatics can be defined as computer programs and analysis methods developed to process big data from telescopes.

For example, CALTECH's space telescope GALEX (The Galaxy Evolution Explorer) 30 TB, Australia's SkyMapper (Southern Sky Survey) 500 Terabyte, and NASA JPL's Hawaii telescope PanSTARRS 40 PB is generating data while the data produced by Palomar Observatory is 3 TB. The amount of data generated reaches almost zettabytes when we combine all the telescopes in the world.

The International Virtual Observatory Association (http://www.ivoa.net) established for this purpose is designed to combine information from telescopes all around the world with Hadoop to establish an environment accessible to every astronomer. A virtual observatory has been set up and all data from telescopes so far has been shared, and when astronomers want to analyze a region, they can access information from telescopes on a single screen and make virtual observations. In short, the virtual observatory makes it easier for scientists to make science.

All observational data in the world and in space were collected and opened to share with the virtual observatory. How will this data be processed? The database mining programs which were developed for them come into play here. CALTECH claims the existence of the ninth planet because it analyzes the data it receives from them and computes mathematics. The main astronomy data analysis programs are:

1. StatCodes (http://astrostatistics.psu.edu/statcodes/)
2. VOStat (http://astrostatistics.psu.edu:8080/vostat/)

3. Weka (http://www.cs.waikato.ac.nz/ml/weka/index.html)
4. AstroML (http://github.com/astroML/astroML)
5. DAME (Data Mining & Exploration) (http://dame.dsf.unina.it/)
6. Auton Lab (http://www.autonlab.org/autonweb/2.html)

These programs and these data alone are not something that astronomers can handle. They need computer engineers, Big Data experts and statisticians to do these tasks. 7 different organizations operate to bring them together. The complete list (ASAIP) is available at: http://asaip.psu.edu.

Data was collected from the telescopes, a virtual observatory was opened, this information was accessed and analyzed with software. The results of these studies was evaluated at an annual conference. The name of this conference is Astronomical Data Analysis Software and Systems. One-year studies have been reviewed and the methods in the analyses compared over the last 6 years at the conference held every year. The website of this conference (ADASS) is available at: http://www.adass.org.

As a result, it can be said that Big Data is a technology that will hold the future in both normal life and astronomy. Programmers are already calling Big Data “Oil of the Future” (Zhang and Zhao, 2015).

3. Astroinformatics

Astroinformatics is an interdisciplinary field of astronomy, astrophysics and informatics that uses information and communications technologies to solve the big data problems in astronomy (Zhang and Zhao, 2015).

In this subheading, let us mention the characteristics of the data scientist who handles the data so we can understand more comprehensively how data becomes information. Then we will relate it with astronomy. An inquisitive mind set enables the data scientist to solve complex problems. Data scientists would normally work in multi-disciplinary teams. This means that you would normally develop an area of expertise and then work in a team to solve problems. There is no one qualification path that will enable you work as a data scientist. For example, some data scientists have a statistics or mathematics academic background, whereas others have combined statistics with computer science and computer programming. Data scientists would often have training in mathematics and statistics, modelling, and computer science and then learn specific technology skills and programming languages to be able to

complete data analysis tasks. There are different approaches to becoming a data scientist and it can be quite confusing once you start reading and talking to people about what matters. The following list of core skills areas is intended as a map to the various skills sets:

1. Fundamentals (including mathematics and data modelling)
2. Statistics (including probability theory, exploratory data analysis, hypothesis testing and regression)
3. Programming (Computer programming languages such as Python, statistical programmes such as R and commercial packages such as SPSS and Hadoop)
4. Machine Learning (knowing which techniques to apply using Python and R)
5. Text Mining/Natural Language Processing (text analysis, packages such as WEKA)
6. Data Visualization (using statistical packages to visualise and present data)
7. Big Data (including Hadoop: it is a free database program and the most widely used framework for distributed file system processing)
8. Data Getting (including data formats, data discovery, and data integration)
9. Data Munging (knowing how to clean data to be able to analyse it)
10. Toolbox (programs and packages that you should be familiar with)

There are specific techniques that have to be learned within each of these areas (https://www.unisa.ac.za/static/corporate_web/Content/About/Service%20departments/DCCD/Documents/career_data_science_math_stats_unisa.pdf).

The Virtual Observatory (VO) became real. The Virtual Observatory is the vision that astronomical datasets and other sources should work within as an uninterrupted whole. Many projects and data centres worldwide are working towards this goal. For example, the International Virtual Observatory Alliance (IVOA) is an organisation that debates and agrees the technical standards that are needed to make the VO possible. Examples of virtual observatory science are:

1. Combine the data from multi-TB, billion-object surveys in the optical, IR, radio, x-ray, etc (to learn about the large-scale structure in the universe and the structure of our galaxy).
2. Discover rare and unusual (one-in-a-million or one-in-a-billion) types of sources (e.g., extremely distant or unusual quasars, new types, etc.).

3. Match Peta-scale numerical simulations of star or galaxy formation with equally large and complex observations.

VO has also been successful because: all data are collected in a digital form, the computer- and data-enthusiast community, some appropriate standard formats, large data collections from funded-agency supported archives, an established culture of data sharing, a community initiative driven by the needs of an exponential data growth, federal agency support/funding, and data have no commercial value or privacy issues.

The positive aspects of the Virtual Observatory are: progress on interoperability, progress on standards etc., a global data grid of astronomy, empowering a broad community, some useful web services, community training, and outreach better than most other fields.

There are also negative aspects to VO: failing on data exploration and mining tools, should this be the level science reaches?, How much can be done effective science with little VO?, a slow community participation, and finally its own end (Djorgovski, 2017). What's better? Answer: Astroinformatics, as it can be bridged from the virtual observatory to astroinformatics as follows:

1. A bridge field connecting astronomy with computer science and engineering.
2. A mechanism for a broader community inclusion both as contributors and as consumers.
3. A mechanism for an interdisciplinary data science methodological sharing with other fields (Djorgovski, 2017).

The fields of Astrostatistics and Astroinformatics are vital for dealing with the actual big data issues in astronomy. The new disciplines of Astrostatistics and Astroinformatics have emerged in order to cope with the various challenges and opportunities offered by the exponential growth of astronomical data volumes, rates, and complexity.

The size of data repositories with the rapid growth of data volume from a variety of sky surveys has increased from gigabytes into terabytes and petabytes. Astroinformatics has appeared at an opportune time to deal with the challenges and opportunities generated by the massive data volume, rates, and complexity from new generation telescopes. This field of study uses data mining tools to analyze large astronomical repositories and surveys. Its many advantages are not only an efficient management of data resources but also the development of new valid tools intended for astronomical problems.

Different scientific areas have similar requirements concerning the ability to handle massive and distributed data sets and to perform complex knowledge discovery tasks on them. Data mining specialists have developed a lot of software and tools for solving various data mining tasks in different fields. Currently, there exist many successful application examples in the fields of business, medicine, science, and engineering. Researchers from astronomy, statistics, informatics, computers, and data mining are collaborating to focus on developing data mining software and tools for use in astronomy. Certainly some data mining tools from other fields may be directly used to overcome astronomical problems.

The arrival of the big data era in astronomy has led to a collaboration boom between astronomers, statisticians, computer scientists, data scientists, and information scientists. Collaboration is the only solution for scientists faced with difficulties and challenges caused by big data. Because of this, various organizations (see Table 2), for example, the International Astro-Statistical Association (IAA, affiliated to the International Statistical Institute), the American Astronomical Society Astroinformatics and Astrostatistics Working Group (AAS/WGAA) were established. Other groups are the Union Working Group in Astro-Statistics and Astroinformatics (IAU/WGAA), the Planned Large Synoptic Survey Telescope (LSST/ISSC) Consortium of Information and Statistics Sciences, the American Society of Statistics in Astroistatistics (ASA/IGA) and the IAA Study Cosmoistatistics Group (Zhang and Zhao, 2015).

4. Big Data in Astronomy

At present, the continuing construction and development of ground-based and space-born sky surveys ranging from gamma rays and x-rays, ultraviolet, optical, and infrared to radio bands is bringing astronomy into the big data era. Astronomical data, already amounting to petabytes, continue to increase with the advent of new instruments. Astronomy, like many other scientific disciplines, is facing a data tsunami that necessitates changes to the means and methodologies used for scientific research. This new era of astronomy is making dramatic improvements in our comprehensive investigations of the Universe. Much progress is being made in the study of such astronomical issues as the nature of dark energy and dark matter, the formation and evolution of galaxies, and the structure of our own Milky Way. Astronomy research is changing from being hypothesis-driven to being data-driven to being data-intensive (Zhang and Zhao, 2015).

What is big-data in astronomy and astrophysics? Some of the big data providers in astronomy (ground and space based telescopes) are given with their abbreviations in Figure 1. They are also presented in tables and other forms.

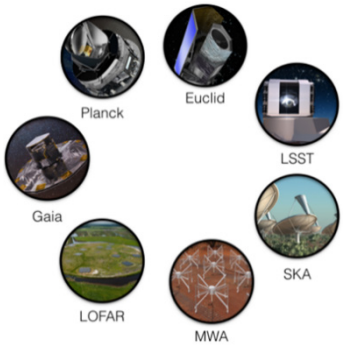

Figure 1: Wide and deep data and observations (McEwen, 2016)

The sky is not the limit for big data! We often hear terms such as "big data" and "data deluge". And it doesn't get much bigger than astronomy and satellite data! (https://adacs.org.au/wp-content/uploads/2018/01/10AstronomyThings.pdf)

Big data cannot be categorized into existing technological dimensions like data mining, algorithms and machine learning, or artificial intelligence. Big data are interconnected with those technologies and takes a new form during this process. As artificial intelligence becomes smarter, more autonomous and opaque, big data are transformed in novel ways. Without big data and the abundance of data available, none of the current improvements in technology would be possible. Big data are entangled in a complex way with data mining, algorithms and machine learning, and artificial intelligence. Big data enable those technologies to be better. On the other hand, big data are enabled by these technologies. Big data contribute to a cycle of technology and can be depicted as in Figure 2.

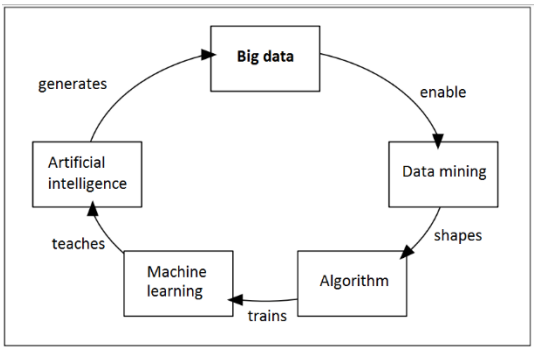

Figure 2: Big Data Technology Cycle (Scholz, 2017)

Researchers are parsing big data produced by the Hubble Space Telescope, the Large Hadron Collider and numerous other sources to learn more about the nature and origins of the

universe. These processes all involve large amounts of information that were once too vast and messy for even computers to analyze. Now that data can be mined for patterns and insights, some of which could spawn major advances in everything from theoretical physics to basic government services. In other words, big data are a chance to take all the things we don't know we know and finally know them (Marks, 2011).

5. Importance of Data in Astronomy

Rapid advances in technology and terabytes of data obtained in a day in astronomy have led to the discovery of new celestial bodies. Thus, in astronomy, data mining applications and algorithms, new decision trees and neurol networks were needed in astronomy for the rapid clustering and classification of these celestial bodies. Table 1 shows data mining tasks and their applications in astronomy (Zhang and Zhao, 2015).

Table 1. Applied approaches as well as their applications for he main data mining tasks in astronomy

Data Mining Tasks	**Applied Approaches**	**Applications in Astronomy**
Classification	Artificial Neural Networks (ANN)	Known knowns: - Spectral classification (stars, galaxies, quasars, supernovae) - Photometric classification (star and galaxies, stars and quazars, supernovae) - Morphological classification of galaxies - Solar activity
	Support Vector Machines (SVM)	
	Learning Vector Quantization (LVQ)	
	Decision Trees	
	Random Forest	
	K-Nearest Neighbors	
	Naive Bayesian Networks	
	Radial Basis Function Network	
	Gaussian Process	
	Decision Table	
	ADTree	
Regression	Artificial Neural Networks (ANN)	Known unkowns: - Photometric redshifts (galaxiex, quazars) - Stellar physical parameter measurement ([Fe/H], T_{eff}, logg)
	Support Vector Regression (SVR)	
	Decision Trees	
	Random Forest	
	K-Nearest Neighbor Regression	
	Kemel Regression	
	Principal Component Regression (PCR)	
	Gaussian Process	
	Least Squared	
	Regression Random	
	Forest	
	Partial Least Squares	

Clustering	Prencipal Component Analysis (PCA)	Unkown unknowns: - Classification - Special/rare object detection
	DBScan	
	K-Means	
	OPTICS	
	Cobweb	
	Self Organizing Map (SOM)	
	Expectation	
	Maximization	
	Hierarchical Clustering	
	AoutuClass	
	Gaussian Mixture Modeling (GMM)	
Outlier Dedection or Anomaly Detection	Prencipal Component Analysis (PCA)	Unkown unknowns: - Special/rare/ object detection
	K-Means	
	Ecpectation	
	Maximization	
	Hierarchical Clustering	
	One-Class SVM	
Time-Series Analysis	Artificial Neural Networks (ANN)	Known unkowns: - Novel detection - Trend prediction
	Support Vector Machines (SVM)	
	Random Forest	

Good organization was needed because ground-based and space-based observations led to collection of very large data. Sky surveys have been made available to astronomers by these organizations. Some of the astrostatistics and astroinformatics organizations are given in Table 2 (Zhang and Zhao, 2015).

Table 2. Astrostatistics and astroinformatics organizations

Organization	**Under community or project**	**Foundation Time**	**Chair**
International Astrostatistics Association (IAA)	The International Statistical Institude (ISI)	August 2012	Joseph Hilbe
IAU Working Group in Astrostatistics and Astroinformatics	The International Astronomical Union (IAU)	August 2012	Eric Feigelson
AAS Working Group in Astroinformatics and Astrostatistics	The American Astronomical Society (AAS)	June 2012	Zeljko Ivezic
ASA Interest Group in Astrostatistics	The American Statistical Association (ASA)	March 2014	Jessi Cisnewski
LSST Informatics and Statistics Science Collaboration	The Large Synoptic Survey Telescope (LSST)	Under construction	Kirk Borne
IAA Working Group on Cosmostatistics (renamed Cosmostatistics initiative, short for COIN)	The International Astrostatistics Association (IAA)	April 2014	Rafael de Souze

These organizations are also available on the ASAIP, Astrostatistics and Astroinformatics Portal of http://asaip.psu.edu. The ASAIP web site will provide links and resources to organizations devoted to the advancement of statistical and computational methodology for

astronomical research. It is intended to promulgate the organizations' work and assist in cross-fertilization between various organizations and interested individuals (Zhang and Zhao, 2015).

The broadest organizations are those associated with international societies:

1. International Astrostatistics Association (IAA): https://asaip.psu.edu/organizations/iaa
2. Special Interest Group in Astrostatistics (https://www.isi-web.org/index.php/news-from-isi/128-isi-astrostatistics-committee-and-network) within The International Statistical Institude (ISI): https://www.isi-web.org
3. Commission on Astroinformatics and Astrostatistics (https://www.iau.org/science/scientific_bodies/commissions/B3/info; under construction) within The International Astronomical Union (IAU): https://www.iau.org
4. The Astrominer Task Force (https://asaip.psu.edu/organizations/ieee-astrominer-task-force) of the Institute of Electrical and Electronics Engineers Computational Intelligence Society (IEEE-CIS): https://cis.ieee.org/data-mining-tc.html

The important U.S. national organizations also use these portions of the ASAIP web site:

1. The Working Group in Astroinformatics and Astrostatistics (https://asaip.psu.edu/organizations/aas-working-group-in-astroinformatics-and-astrostatistics) within the American Americal Society (AAS): https://aas.org
2. The Interest Group in Astrostatictics (https://asaip.psu.edu/organizations/asa-interest-group-in-astrostatistics) within the International Astrostatistics Association (IAA): https://www.amstat.org

One Project-level organization is using this web site:

1. The Informatics and Statistics Science Collaboration (https://asaip.psu.edu/organizations/lsst-informatics-and-statistics-Science-collaboration) of Large Synoptic Survey Telescope (LSST): https://www.lsst.org
2. International Astrostatistics Association (IAA): https://asaip.psu.edu/organizations/iaa
3. ISI Astrostatistics Special Interest Group: https://asaip.psu.edu/organizations/isi-astrostatistics-special-interest-group-sigastro

4. IAU Commission on Astroinformatics and Astrostatistics: https://asaip.psu.edu/organizations/iau-commission-on-astroinformatics-and-astrostatistics
5. IEE CIS Task Force on Mining ComplexAstronomical Data: https://asaip.psu.edu/organizations/ieee-astrominer-task-force
6. AAS Working Group in Astroinformatics and Astrostatistics: https://asaip.psu.edu/organizations/aas-working-group-in-astroinformatics-and-astrostatistics
7. ASA Interest Group in Astrostatistics: https://asaip.psu.edu/organizations/asa-interest-group-in-astrostatistics
8. LSST Informatics and Statistics Science Collaboration: https://asaip.psu.edu/organizations/lsst-information-and-statistical-science-collaboration
9. The Virtual Observatory (VO): http://www.ivoa.net; http://www.euro-vo.org; https://heasarc.gsfc.nasa.gov/vo/summary
10. The All Sky Virtual Observatory (ASVO): http://www.asvo.org.au/

The ASAIP provides searchable abstracts to recent papers in the field, several discussion forums, various resourches for research, brief articles by experts, lists of meetings, and access to various web resources such as on-line courses, books, jobs and blogs.

6. Data Sources in Astronomy

There are many sky surveys for ground-based and space-based observations at different wavelengths of the electromagnetic spectrum. The most important ones are given in Table 3 according to the data volumes (added to the table given by Zhang and Zhao, 2015). Similarly, foreseen data measurements of similar or different surveys are given in Figure 3. Kremer et al. (2017) provide detailed information about the data size of some large space surveys obtained overnight (see Figure 4). The data growth in only one area, for example the radio region, is given in Figure 5.

Table 3. Data volumes of different sky survey projects	
Sky Survey Projects	**Data Volume**
Very Large Telescope (VLT, per night in 1998)	10 GB
The Hubble Space Telescope (HST; per week)	20 GB
Visible and Infrared Telescope for Astronomy (VISTA, per night in 2009)	315 GB
GAIA satellite (since 2014.07.25)	69585 GB
The Polamar Digital Sky Survey (DPOSS)	3 TB
The Two Micron All-Sky Survey (2MASS)	10 TB
Green Bank Telescope (GBT)	20 TB
The Galaxy Evolution Explorer	30 TB
The Sloan Digital Sky Survey (SDSS; 200 GB per night in 2000)	40 TB
Thirty Meter Telescope (TMT; per night in 2022)	90 TB
Sky Mapper Southern Sky Survey	500 TB
The Panoramic Survey Telescope and Rapid Responce System (PanSTARRS)	~40 PB expected
Cherenkov Telescope Array (CTA; by 2030)	~100 PB expected
The Square Kilometre Array (SKA Observatory; 500 and 1000 GB per second for low and mid, respectively)	~300 PB expected
The Large Synoptic Survey Telescope (LSST; ~2 TB per hour, 15-30 TB per night)	~4.6 EB expected

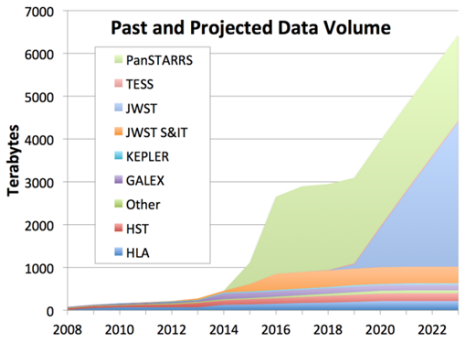

Figure 3: Past and projected growth in the data volume hosted by STScI's Mikulski Archive for Space Telescopes (MAST). In 2016, MAST's holdings exceeded 2.5 Petabytes (https://archive.stsci.edu/reports/BigDataSDTReport_Final.pdf)

.

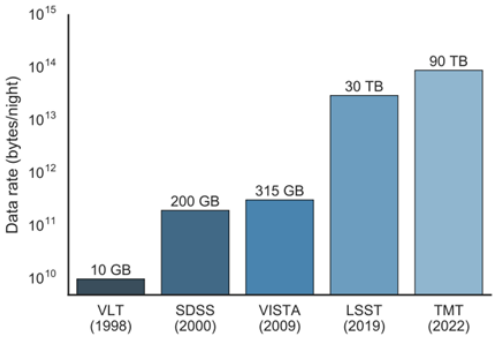

Figure 4: Increasing data volumes of existing and upcoming telescopes: Very Large Telescope (VLT), Sloan Digital Sky Survey (SDSS), Visible and Infrared Telescope for Astronomy (VISTA), Large Synoptic Survey Telescope (LSST) and Thirty Meter Telescope (TMT) (Kremer et al., 2017)

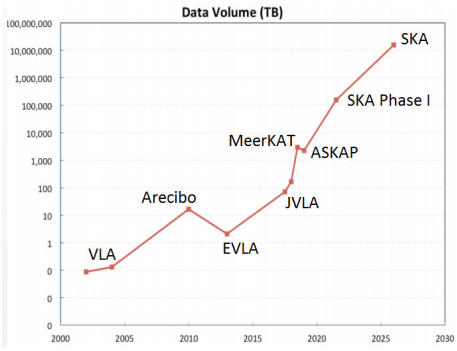

Figure 5: Exponential growth of radio data volumes. The horizontal axis is the year and the vertical is Terabayt (Raynard, 2017)

For developing data-rich astronomy it can be said that telescope + instrument is just a front end for data systems of live actions. Another example of a Big Data science driven by the advances in computing/information technology is presented in Figure 6.

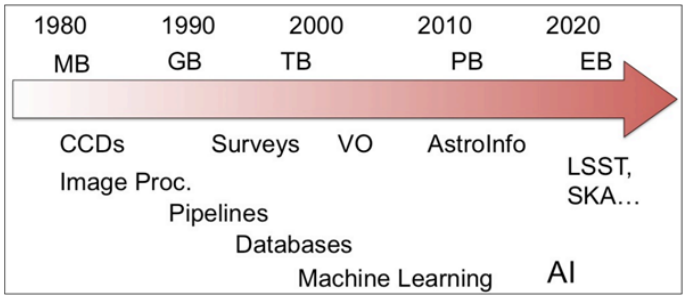

Figure 6: Evolving data-rich astronomy (Djorgovski, 2017)

Given the amount of data in the future archives, we expect that server-side analyses will be commonplace for the users, thus an advanced scripting capability must be supported (Big Data @ STScI). Accordingly, taking into account the number of tiles, years, and filters, the disk space needs for each survey can then be estimated (see Table 4; https://archive.stsci.edu/reports/BigDataSDTReport_Final.pdf). Not surprisingly, for the surveys with large plate scales (PTF, ASASSN, ATLAS), the disk space needs for the stacks and difference images are small and on the order of a few Terabytes (TB). For the surveys with small plate scales (PS1 and LSST), the stacks and difference images are on the order of 1 Petabyte (PB), which goes beyond simple local storage systems, but is feasible on department or campus-wide computing centers. The input data volume in Table 4 seems to be significantly larger. For PTF and ASASSN, it is still feasible to store the input images locally. However, for ATLAS, PS1, and LSST, the input image set is on the order of a PB (ATLAS, PS1) and 17 PBs (LSST) due to the many epochs and/or small plate scales. In these cases, most likely the input images need to be accessed as needed via the Internet. In particular for LSST, this requires excellent connectivity to ensure the data transfer, data reduction and data analysis can be completed on timescales of just a few months (Morgan, 2018).

Table 4. Disk space and computing time requirements for the different surveys (sorted from top to bottom according to input data in the first column)

Survey Name	Input Image Data Volume (TB)	Stacked Image Data Volume (TB)	Stacked Image Processing Time (CPU Days)	Difference Image Data Volume (TB)	Difference Image Processing Time (CPU Days)
LSST	17,107	855	21,094	770	18,984
PanSTARRS	876	219	1080	164	4,050
ATLAS	475	0.8	2,344	0.5	52
PTF	51	3.4	253	2.6	253
ASASSN	6.8	0.04	33	0.03	3

In addition, there are many astronomical archives on the internet about the published articles in astronomy and the properties of celestial bodies. The information about the telescopes mentioned here can be learned from their related sites and was not also written in order not to convert this chapter into the basic astronomy. Some of these are:

1. The Sloan Digital Sky Survey: https://www.sdss.org
2. The Very Large Telescope array (VLT): https://www.eso.org/public/teles-instr/paranal-observatory/vlt
3. Thirty Meter Telescope (TMT): https://www.tmt.org
4. Square Kilometre Array (SKA): https://www.skatelescope.org

5. James Webb Space Telescope (JWST): https://www.jwst.nasa.gov
6. EUCLID Space Telescope: http://www.euclid-ec.org
7. PLANCK satellite: http://www.esa.int/planck
8. The International Event Horizon Telescope (EHT): https://eventhorizontelescope.org
9. Transiting Exoplanet Survey Satallite (TESS): https://www.nasa.gov/tess-transiting-exoplanet-survey-satellite
10. The Murchison Widefield Array (MWA): http://www.mwatelescope.org
11. The Galactic and Extra-Galactıc All-Sky MWA Survey (GLEAM): http://www.mwatelescope.org/gleam
12. The Galactıc and Extra-Galactic All-Sky MWA Extended Survey (GLEAM-X): http://www.mwatelescope.org/gleam-x
13. The Low-Frequency Array (LOFAR): http://www.lofar.org
14. The Solar & Heliospheric Observatory (SOHO): https://sohowww.nascom.nasa.gov
15. Sunspotter: https://www.sunspotter.org
16. Visible and Infrared Survey Telescope for Astronomy (VISTA): http://www.vista.ac.uk
17. Aladin Sky Atlas: https://aladin.u-strasbg.fr
18. SIMBAD Astronomical Database – CDS (Strasbourg): http://simbad.u-strasbg.fr/simbad
19. VizieR: http://vizier.cfa.harvard.edu
20. SAO/NASA ADS, Astrophysics Data System (ADS): http://cdsads.u-strasbg.fr
21. Minor Planet Center (MPC): https://minorplanetcenter.net/iau/mpc.html
22. SETI Institute: https://www.seti.org
23. The Galaxy Zoo website: https://www.galaxyzoo.org ; https://www.zooniverse.org/projects/zookeeper/galaxy-zoo
24. Much of the Kepler data for exoplanet discovery is publicly available through Mikulski Archive for Space Telescopes: http://archive.stsci.edu/kepler
25. Kepler Spacecraft: https://www.nasa.gov/mission_pages/kepler/main/index.html
26. Debrecen Sunspot Data archive: http://fenyi.solarobs.unideb.hu/ESA/HMIDD.html

The main purpose of the Gaia (http://sci.esa.int/gaia) in Table 3 is to examine our galaxy and star contents and to provide high precision astrometric and photometric parameters. The satellite will also conduct many other investigations, in particular observing millions of small galaxies of the local universe, a significant number of supernovae, and interactive binary systems as a large part of transient events.

The Gaia satellite (https://www.cosmos.esa.int/web/gaia/mission-numbers) continues to provide precise location data. Continuity in the acquisition of observational data within the framework of Gaia campaigns with ground-based telescopes has provided and so increased the performance of the data. A recent example of this was the binary microlensing event of the Gaia16eye object (Wyrzykowski et al., 2020). Gaia16aye is the first of its kind, the first activity discovered by the Gaia space mission and discovered in the direction of the Northern Galactic Disc. The light curve exhibited five different maximum brightnesses of up to 11 magnitudes, and the event was elaborated with approximately 25000 data points collected by the telescope network organized by the Gaia team for 500 days. This study demonstrated the potential of the microlensing method to question the mass function of dark objects, including black holes, in other directions from the Galactic overhang. This also emphasizes the importance of long-term coordinated observations with a network of heterogeneous telescopes.

The observation of all asteroids by a single observatory is not possible because their number, being more than a million, is too big to handle. For this reason, it is necessary that all astronomers and sky enthusiasts on Earth should work together. All observations of asteroids, comets and natural satellite observations of the Solar System by observers or amateur astronomers are collected in the Minor Planet Center (MPC; https://minorplanetcenter.net/iau/mpc.html). The ephemeris information and the assigned trajectory parameters are published on the web address which is open to all internet users worldwide. At the same time the accuracy of the parameters obtained can be controlled by anyone interested in this field (Kaynar, 2019).

7. Big Data for Explorations of the Universe

Machine Learning techniques have begun to find applications in astronomy, but mainly for “clerical” tasks, such as error checking, and bulk classification. This leaves vast scope for harnessing Machine Learning for more interesting tasks that enable new scientific discoveries. Historically, major discoveries have often relied on serendipity; an expert examines new data with an eagle eye and an open mind. However effective this approach has been in the past, it does not scale. New astronomical datasets are too massive and complex for any individual or group of experts to look at every aspect, object, or measurement. Yet modern algorithms for

sequencing, classification, or anomaly detection can provide us with methods to uncover new phenomena (Mazeh & Poznanski, 2018).

Within the next few years, image analysis and machine learning systems that can process terabytes of data in near real-time with high accuracy will be essential (Gómez-Vargas, 2018). There are great opportunities for making novel discoveries, even in databases that have been available for decades. The volunteers of Galaxy Zoo have demonstrated this multiple times by discovering structures in SDSS images that have later been confirmed to be new types of objects. These volunteers are not trained scientists, yet they make new scientific discoveries. Even today, only a fraction of the images of SDSS have been inspected by 12 humans. Without doubt, the data still hold many surprises, and upcoming surveys, such as LSST, are bound to image previously unknown objects. It will not be possible to manually inspect all images produced by these surveys, making advanced image analysis and machine learning algorithms of vital importance. One may use such systems to answer questions like how many types of galaxies there are, what distinguishes the different classes, whether the current classification scheme is good enough, and whether there are important sub-classes or undiscovered classes. These questions require data science knowledge rather than astrophysical knowledge, yet the discoveries will still help astrophysics tremendously. In this new data-rich era, astronomy and computer science can benefit greatly from each other. There are new problems to be tackled, novel discoveries to be made, and above all, new knowledge to be gained in both fields (Kremer et al., 2017).

Big Data is transforming how astronomers make discoveries. The next game-changer is likely lurking in the data we already have but it will take scientists years to uncover it. Earlier in 2018, astronomers stumbled upon a fascinating finding: Thousands of black holes likely exist near the center of our galaxy. The x-ray images that enabled this discovery weren't from some state-of-the-art new telescope. Nor were they even recently taken - some of the data was collected nearly 20 years ago. The researchers discovered the black holes by digging through old, long-archived data. Discoveries like this will only become more common, as the era of "big data" changes how science is done. Astronomers are gathering an exponentially greater amount of data every day so much that it will take years to uncover all the hidden signals buried in the archives. Sixty years ago, the typical astronomer worked largely alone or in a small team. They likely had access to a respectably large ground-based optical telescope at their home institution. Their observations were largely confined to optical wavelengths more or less what the eye can see. That meant they missed signals from a host of astrophysical sources, which can emit non-visible radiation from very low-frequency radio

all the way up to high energy gamma-rays. For the most part, if you wanted to do astronomy, you had to be an academic or eccentric rich person with access to a good telescope. Old data were stored in the form of photographic plates or published catalogs. But accessing archives from other observatories could be difficult and it was virtually impossible for amateur astronomers. Today, there are observatories that cover the entire electromagnetic spectrum. No longer operated by single institutions, these state-of-the-art observatories are usually launched by space agencies and are often joint efforts involving many countries. With the coming of the digital age, almost all data are publicly available shortly after it is obtained. This makes astronomy very democratic - anyone who wants to can reanalyze almost any data set that makes the news (you too can look at the Chandra data that led to the discovery of thousands of black holes!). These observatories generate a staggering amount of data. For example, the Hubble Space Telescope (HST), operating since 1990, has made over 1.3 million observations and transmits around 20 GB of raw data every week, which is impressive for a telescope first designed in the 1970s. The Atacama Large Millimeter Array (ALMA) in Chile now anticipates adding 2 TB of data to its archives every day. The archives of astronomical data are already impressively large. But things are about to explode. Each generation of observatories is usually at least 10 times more sensitive than the previous, either because of improved technology or because the mission is simply larger. Depending on how long a new mission runs, it can detect hundreds of times more astronomical sources than previous missions at that wavelength. For example, compare the early EGRET gamma ray observatory, which flew in the 1990s, to NASA's flagship mission Fermi, which turns 10 in 2018. EGRET detected only about 190 gamma ray sources in the sky. Fermi has seen over 5,000. The Large Synoptic Survey Telescope (LSST), an optical telescope currently under construction in Chile, will image the entire sky every few nights (Estévez, 2016). It will be so sensitive that it will generate 10 million alerts per night on new or transient sources, leading to a catalog of over 15 petabytes after 10 years. The Square Kilometre Array (SKA), when completed in 2020, will be the most sensitive telescope in the world, capable of detecting airport radar stations of alien civilizations up to 50 light-years away (Scaife, 2016 and 2019). In just one year of activity, it will generate more data than the entire internet. These ambitious projects will test scientists' ability to handle data. Images will need to be automatically processed meaning that the data will need to be reduced down to a manageable size or transformed into a finished product. The new observatories are pushing the envelope of computational power, requiring facilities capable of processing hundreds of terabytes per day. The resulting archives all publicly searchable will contain 1 million times more information that what can be stored on your typical 1 TB backup disk. The data deluge will

make astronomy become a more collaborative and open science than ever before. Thanks to internet archives, robust learning communities and new outreach initiatives, citizens can now participate in science. For example, with the computer program Einstein@Home, anyone can use their computer's idle time to help search for gravitational waves from colliding black holes. It's an exciting time for scientists, too. Astronomers often study physical phenomena on timescales so wildly beyond the typical human lifetime that watching them in real-time just isn't going to happen. Events like a typical galaxy merger (which is exactly what it sounds like), can take hundreds of millions of years. All we can capture is a snapshot, like a single still frame from a video of a car accident. However, there are some phenomena that occur on shorter timescales, taking just a few decades, years or even seconds. That's how scientists discovered those thousands of black holes in the new study. It's also how they recently realized that the x-ray emission from the center of a nearby dwarf galaxy has been fading since first detected in the 1990s. These new discoveries suggest that more will be found in archival data spanning decades. In Meyer's (2018) work, she used Hubble archives to make movies of "jets" of high-speed plasma ejected in beams from black holes. She used over 400 raw images spanning 13 years to make a movie of the jet in nearby galaxy M87. That movie showed, for the first time, the twisting motions of the plasma, suggesting that the jet has a helical structure. This kind of work was only possible because other observers, for other purposes, just happened to capture images of the source she was interested in, back when she was in kindergarten. As astronomical images become larger, higher resolution and ever more sensitive, this kind of research will become the norm (Meyer, 2018).

8. Ways of Processing Big Data in Astronomy

Now and the next decade promises to be an exciting time for astronomers. Large volumes of astronomical data are continuously being collected from highly productive space missions. These data have to be efficiently stored and analyzed in such a way that astronomers maximize their scientific return from these missions. Recognizing the need to better handle astronomical datasets, we designed ASTROIDE, a distributed data server for astronomical data. We analyze the peculiarities of the data and the queries in cosmological applications and design a new framework where astronomers can explore and manage vast amounts of data. ASTROIDE introduces effective methods for efficient astronomical query execution on Spark through data partitioning with HEALPix and customized optimizer. ASTROIDE offers a simple, expressive and unified interface through ADQL, a standard language for querying databases in astronomy. Experiments have shown that ASTROIDE is effective in processing astronomical data, scalable and outperforms the state-of-the-art (Brahem et al., 2018).

Mehta et al. (2017) presented the first comprehensive study of large-scale image analytics on big data systems. They surveyed the different paradigms of large-scale data processing platforms using two real-world use cases from domain sciences. While they could execute the use cases on these systems, their analysis shows that leveraging the benefits of all systems requires deep technical expertise. Overall, they argue that current systems provide good support for image analytics, but they also open new opportunities for further improvement and future research.

Zhang et al. (2016) investigated the idea of leveraging the modern big data platform for many-task scientific applications. Specifically, they built Kira (https://github.com/BIDS/Kira), a flexible, scalable, and performant astronomy image processing toolkit using Apache Spark running on Amazon EC2 Cloud. They also presented the real world Kira Source Extractor application, and use this application to study the programming flexibility, dataflow richness, scheduling capacity and performance of the surrounding ecosystem. They also demonstrated that Apache Spark can integrate with a pre-existing astronomy image processing library. This allows users to reuse existing source code to build new analysis pipelines. They believe that Apache Spark's flexible programming interface, rich dataflow support, task scheduling capacity, locality optimization, and built-in support for fault tolerance make Apache Spark a strong candidate to support many-task scientific applications. Apache Spark is one (popular) example of a Big Data platform. They learned that leveraging such a platform would enable scientists to benefit from the rapid pace of innovation and large range of systems and technologies that are being driven by widespread interest in Big Data analytics. Their experience with Kira demonstrates that data intensive computing platforms like Apache Spark are a performant alternative for many-task scientific applications.

Examples of other special methods for analyzing big data include: Bayesian analysis, MCMC sampling, hierarchical probabilistic (Bayesian) models, variable selection, experimental design, machine learning, optimisation, wavelets, sparsity, compressed sensing, and finally Astrostatics and Astroinformatics.

We conclude the end of this chapter of the book with this last sentence: collaborations between astronomers, statisticians and information scientists have begun, but need to be expanded. The International Statistical Institute and similar astronomical organisations should be promoting to continue these collaborations (Feigelson & Babu, 2012). Bigger data is not always better data and may big data be with you (Scholz, 2017).

9. Conclusion

Thanks to the big data in astronomy and Machine Learning algorithms, there have been great advances in Astroinformatics and Astrostatistics studies. Very valuable information has been obtained about micro and macro structures of the universe. Wide horizons have been opened to astronomers about dark matter, dark energy, supernovae, novae and galaxies. It is tried to develop algorithms that can make predictions about solar eruptions by observing the atmosphere of our sun continuously. In addition, the research continues with large data following the trajectory movements of asteroids threatening our world. Through surveys such as GAIA, LSST and TMT, there will be many new developments and discoveries in Physics, Astrophysics, Astronomy, and Cosmology.

In this chapter we tried to guide those who want to study Astronomy, Astroinformatics and Astrostatistics. We can say that the field of astronomy is at the top of the big data in the world. We hope the references section of this chapter will guide enthusiasts. Everything about big data in astronomy is almost impossible to write here. For this reason, the article titled "the Astro2020 Science White Paper. The Next Decate of Astroinformatics and Astrostatistic" prepared by Aneta Siemiginowska (2019) together with 34 authors is a good reference for those who wish to work in these fields.

Acknowledgments. We thank Demir IT Company (Eskisehir, Turkey) for providing computer support. HHE thanks TUBITAK National Observatory (TUG) for the experience from several projects on follow-up observations of the Gaia satellite since started scientific operations in mid-2014.

References

Brahem, M., Yeh, L., Zeitouni, K. (2018). *ASTROIDE: A Unified Astronomical Big Data Processing Engine over Spark*. A Preprint, October 25.

Dindar, M., Helhel, S., Esenoglu, H., Parmaksizoglu, M. (2015). A new software on TUG-T60 autonomous telescope for astronomical transient events. *Experimental Astronomy*, 39(1), 21–28

Djorgovski, S., G. (2017). *Astronomy in the Era of Big Data- From Virtual Observatory to Astroinformatics and beyond*. TIARA Summer School on Astrostatistics and Big Data Taipei, Taiwan, September.

Gómez-Vargas, G. (2018). *First Ideas to Connect Astronomical Data, Deep Learning and Image Analysis, Accelerating the search of dark matter with machine learning*. Lorentz Center, Leiden, January.

Estévez, P. (2016). *Big Data Era Challenges and Opportunities in Astronomy: How SOM/LVQ and Related Learning Methods Can Contribute?* WSOM 2016 Houston, TX, January 8.

Feigelson, E. D. & Babu, G. J. (2012). Big data in astronomy. *Significance*, The Royal Statistical Society, August.

Kaynar, S. (2019). *Determination of the Trajectory of Selected Several Near Earth Asteroids and Investigation Their Physical Properties*. Akdeniz University Graduate School of Natural and Applied Sciences Department of Physics Master Thesis (October).

Kremer, J., Kristoffer, S. S., Gieseke, F., Steenstrup, K. P., Igel, C. (2017). Big Universe, Big Data: Machine Learning and Image Analysis for Astronomy, *IEEE Intelligent Systems*, 32, 16–22, March–April (https://arxiv.org/abs/1704.04650).

Marks, J. (2011). *5 Things You Need to Know About Big Data*. NetApp. Veteran Data Solutions-VetDS

Mazeh, T. & Poznanski, D. (2018). *Big Data and Exo-Planets*, Proposal for a research group in Astronomy.

McEwen, J. (2016). *Big-Data in Astronomy and Astrophysics Extracting Meaning from Big-Data*.

(https://indico.hephy.oeaw.ac.at/event/86/session/3/contribution/1/material/slides/0.pdf)

Mehta, P., Dorkenwald, S., Zhao, D., Kaftan, T., Cheung, A., Balazinska, M., Rokem, A. (2017). *Comparative Evaluation of Big-Data Systems on Scientific Image Analytics Workloads*. Andrew Connolly, Jacob Vanderplas, Yusra AlSayyad, Proceedings of the VLDB Endowment, Vol. 10, No. 11.

Meyer, E. (2018). Big Data is Transforming How Astronomers Make Discoveries, *The Conversation*, May 15 (https://theconversation.com/the-next-big-discovery-in-astronomy-scientists-probably-found-it-years-ago-but-they-dont-know-it-yet-95280?xid=PS_smithsonian)

Morgan, H. (2018). *Large Synoptic Survey Telescope (LSST) Scaling Issues and Network Needs*, Pacific Northwest Gigapop Meeting October 23.

Raynard, L. (2017). *Radio Astronomy & SDGs A Justification or Solution?* South African Radio Astronomy Observatury (SARAO), September 4.

Scaife, A. (2019). *Big Telescope, Big Data: Towards Exa-Scale With the SKA, Numerical algorithms for high-performance computational science*. Royal Society 8–9 April.

Scaife, A. (2016). *Big Telescope, Big Data: Indirect Imaging in the SKA Era*, IAU Astroinformatics, Sorrento.

Scholz, T. M. (2017). *Big data in organizations and the role of human resource management: A complex systems theorybased conceptualization*. Econstor, Personalmanagement und Organisation, No. 5, Peter Lang International Academic Publishers (http://hdl.handle.net/10419/182489)

Siemiginowska, A., Eadie, G., Czekala, I. with 33 authors. (2019). *Astro2020 Science White Paper: The Next Decade of Astroinformatics and Astrostatistics*. 15 March. (https://arxiv.org/abs/1903.06796)

Wyrzykowski, L. et al. (2020). Full Orbital Solution for the Binary System in the Northern Galactic Disk Microlensing Event Gaia16aye, *Astronomy & Astrophysics*, in pressed (633, A98.)

Zhang, Y., Zhao, Y. (2015). Astronomy in the Big Data Era. *Data Science Journal*, 14(11), 1–9 (https://datascience.codata.org/articles/10.5334/dsj-2015-011).

Zhang, Z., Barbary, K., Nothaft, F. A., Sparks, E. R., Zahn, O., Franklin, M. J., Patterson, D. A., Perlmutter, S. (2016). Kira: Processing Astronomy Imagery Using Big Data Technology, *IEEE Transactions on Big Data*, DOI: 10.1109/TBDATA.2016.2599926.