

PROCEEDINGS PAPER

Astronomy in the Big Data Era

Yanxia Zhang¹ and Yongheng Zhao¹

¹ Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, 20A Datun Road, Chaoyang District, 100012, Beijing, China
zyx@bao.ac.cn, zyx@lamos.org

The fields of Astrostatistics and Astroinformatics are vital for dealing with the big data issues now faced by astronomy. Like other disciplines in the big data era, astronomy has many V characteristics. In this paper, we list the different data mining algorithms used in astronomy, along with data mining software and tools related to astronomical applications. We present SDSS, a project often referred to by other astronomical projects, as the most successful sky survey in the history of astronomy and describe the factors influencing its success. We also discuss the success of Astrostatistics and Astroinformatics organizations and the conferences and summer schools on these issues that are held annually. All the above indicates that astronomers and scientists from other areas are ready to face the challenges and opportunities provided by massive data volume.

Keywords: Big data; Data mining; Astrostatistics; Astroinformatics

1 Introduction

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. To cope with the various challenges and opportunities offered by the exponential growth of astronomical data volumes, rates, and complexity, the new disciplines of Astrostatistics and Astroinformatics have emerged. The following are simple definitions related to astronomy.

Astronomy is the study of the physics, chemistry, and evolution of celestial objects and phenomena that originate outside the Earth's atmosphere, including supernovae explosions, gamma ray bursts, and cosmic microwave background radiation.

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, statistical mechanics, thermodynamics, quantum mechanics, relativity, nuclear and particle physics, and atomic and molecular physics to solve astronomical issues.

Astrostatistics is an interdisciplinary field of astronomy/astrophysics and statistics that applies statistics to the study and analysis of astronomical data.

Astroinformatics is an interdisciplinary field of astronomy/astrophysics and informatics that uses information/communications technologies to solve the big data problems faced in astronomy.

2 Big Data in Astronomy

Different researchers give different definitions of the characteristics of big data. Kirk Borne put forward “10V”: Volume, Variety, Velocity, Veracity, Validity, Value, Variability, Venue, Vocabulary, and Vagueness

(<http://www.mapr.com/blog/top-10-big-data-challenges-%E2%80%93-serious-look-10-big-data-%E2%80%99s>). In our studies, however, we pay more attention to the four Vs of astronomical data: Volume, Variety, Velocity, and Value.

Volume is the amount of data. Data are measured by terabytes, petabytes, and even exabytes. Thus big data pose challenges for capture, cleaning, curation, integration, storage, processing, indexing, search, sharing, transferring, mining, analysis, and visualization. Traditional tools cannot deal with such large amounts of data. Various ground- and space-based large sky survey projects bring a data avalanche into all aspects of astronomy. The data volumes generated by different sky surveys are shown in **Table 1**.

Sky Survey Projects	Data Volume
DPOSS (The Palomar Digital Sky Survey)	3 TB
2MASS (The Two Micron All-Sky Survey)	10 TB
GBT (Green Bank Telescope)	20 PB
GALEX (The Galaxy Evolution Explorer)	30 TB
SDSS (The Sloan Digital Sky Survey)	40 TB
SkyMapper Southern Sky Survey	500 TB
PanSTARRS (The Panoramic Survey Telescope and Rapid Response System)	~ 40 PB expected
LSST (The Large Synoptic Survey Telescope)	~ 200 PB expected
SKA (The Square Kilometer Array)	~ 4.6 EB expected

Table 1: Data volumes of different sky survey projects.

Variety points to data complexity. Astronomical data mainly include images, spectra, time-series data, and simulation data. Most of the data are saved in catalogues or databases. The data from different telescopes or projects have their own formats, which causes difficulty with integrating data from various sources in the analysis phase. In general, each data item has a thousand or more features; this causes a large dimensionality problem. Moreover, data have many data types: structured, semi-structured, unstructured, and mixed.

Velocity means the speed of producing, transmitting, and analyzing data. Speaking of data volume, LSST will generate one SDSS each night for 10 years. Obviously, batch, streams, near-time, or real-time data analysis are necessary. LSST expects it will find 1000 new supernovae each night for 10 years, which suggests that at least 10-100 thousand alerts will be requested. How to efficiently mine, correctly classify, and target the supernovae candidates and make follow-up observations in 10 years time is a huge challenge for astronomers.

Value characterizes the high value to astronomy of the data. It is interesting and inspiring in astronomy to discover surprising, rare, unexpected, and new objects or phenomena. Similarly the discovery of a new distribution trend or law is of great value.

3 Virtual Observatory

Szalay and Gray (2001) first put forward the concept of the Virtual Observatory. The virtual observatory (VO) (http://en.wikipedia.org/wiki/Virtual_observatory) is a collection of interoperating data archives and software tools that utilize the internet to form a scientific research environment in which astronomical research programs can be conducted. The main goal of the VO is to provide transparent and distributed access to data with worldwide availability, which helps scientists to discover, access, analyze, and combine nature and lab data from heterogeneous data collections in a user-friendly manner. The International Virtual Observatory Alliance (IVOA) (<http://www.ivoa.net/>) is an organization that formulates the technical standards that make the VO possible. It also acts as a focus for VO aspirations, a framework for discussing and sharing VO ideas and technology, and a body for promoting and publicizing the VO.

From the deployer's point of view, the VO (http://www.ivoa.net/deployers/intro_to_vo_concepts.html) is not a concrete item like a data warehouse. Rather, it is more like an ecosystem of mutually compatible datasets, resources, services, and software tools that use a common set of technologies and a common set of standards. The idea is to make all these things interoperable - i.e., to make them work nicely together. However, the VO is more than just a set of rules for everyone to follow; it also requires some specialized middleware to glue things together, for example, registry services, distributed storage, sign-on services, and so on. From a more user-centered explanation, using the VO is really just a question of getting familiar with tools and data services that understand VO rules. Up to now, the VO has developed a wide range of services/tools, see the website <http://www.ivoa.net/astronomers/applications.html>. If scientists want to use VO services, they login into <http://www.ivoa.net/>, which introduces VO and explains how to use the VO services/tools, which the scientists then choose according to their requirements. The VO will dramatically improve our ability to do astronomical research that integrates data from multiple instruments. The VO is also a wonderful platform for teaching astronomy, scientific discovery, and computational science. In conclusion, the VO allows scientists to do much more science more easily. There are a growing number of papers being published about VO services/tools.

4 Data Mining

With the rapid growth of data volume from a variety of sky surveys, the size of data repositories 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 next-generation telescopes. This field of study uses data mining tools to analyze large astronomical repositories and surveys. Its key advantages are not only an efficient management of data resources but also the development of new valid tools that address astronomical problems.

Data mining is of great importance in the big data era. It helps researchers to efficiently and effectively discover potential and useful information or knowledge from the large amounts of data that are stored in databases, data warehouses, and other information repositories for data management, analysis, and decision support. According to the type of patterns being mined, data mining tasks mainly consist of summarization, classification, regression, clustering, association, time-series analysis, and outlier/anomaly detection. There have been many reviews of the use of data mining in astronomy (Zhang, Zhao, & Cui, 2002; Zhang, Zheng, & Zhao, 2008; Borne, 2009; Ball, & Brunner, 2010; Zhang & Zhao, 2011). Other reviews include the application of neural networks in astronomy (Tagliaferri, Longo Milano, et al., 2003; Li, Zhang, Zhao, Yang, 2006) and outlier detection in astronomical data (Zhang, Luo, & Zhao, 2004). In summary, **Table 2** shows the approaches and applications most often used in astronomy to do major data mining tasks. Larger scale, deeper, multi-wavelength, and time domain sky surveys lead to a dimensional increase in astronomical data while high-dimensional data cause the curse of dimensionality and inefficient operation or inoperation of many algorithms. In order to improve the efficiency and effectiveness of data mining approaches, feature selection/extraction is necessary when preparing the data. Feature selection is preferred to feature extraction because the former keeps the physical attributes of objects while the latter loses the meaning of the features. **Table 3** gives some often used feature selection/extraction methods. Zheng and Zhang (2008) and references in their paper discuss feature selection/extraction methods. Many books focusing on data mining in astronomy have been written. For example, Murtagh and Heck (1987) compiled "Multivariate Data Analysis", which introduces multivariate data analysis within astronomy and provides these algorithms in Fortran, C, and Java codes. Wall and Jenkins (2003) wrote a book "Practical Statistics for Astronomers", which is a practical handbook on statistics in astronomy. Feigelson and Babu (2012) published a book titled "Modern Statistical Methods for Astronomy with R Applications", which introduces the use of R to analyze astronomical data. The book "Advances in Machine Learning and Data Mining for Astronomy", edited by Way, Scargle, Ali, and Srivastava (2012), reviewed various data mining tools and techniques used by astronomers. The book "Statistics, Data Mining, and Machine Learning in Astronomy", written by Ivezić, Connolly, VanderPlas, and Gray (2014), is a practical python guide for the analysis of survey data. Edwards and Gaber (2014) wrote a book titled "Astronomy and Big Data", which describes a data clustering approach to identifying uncertain galaxy morphology. Starck and Murtagh (2006) published a book "Astronomical Image and Data Analysis" (second edition) about data analysis of astronomical images, and Starck, Murtagh, and Fadili (2010) wrote "Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity".

Data Mining Tasks	Applied Approaches	Applications in Astronomy
Classification	Artificial Neural Networks (ANN) Support Vector Machines (SVM) Learning Vector Quantization (LVQ) Decision Trees Random Forest K-Nearest Neighbors Naïve Bayesian Networks Radial Basis Function Network Gaussian Process Decision Table ADTree	Known knowns: – Spectral classification (stars, galaxies, quasars, supernovas) – Photometric classification (stars and galaxies, stars and quasars, supernovas) – Morphological classification of galaxies – Solar activity
Regression	Artificial Neural Networks (ANN) Support Vector Regression (SVR) Decision Trees Random Forest K-Nearest Neighbor Regression Kernel Regression Principal Component Regression (PCR) Gaussian Process Least Squared Regression Random Forest Partial Least Squares	Known unknowns: – Photometric redshifts (galaxies, quasars) – Stellar physical parameter measurement ([Fe/H], Teff, logg)
Clustering	Principal Component Analysis (PCA) DBScan K-Means OPTICS Cobweb Self Organizing Map (SOM) Expectation Maximization Hierarchical Clustering AutoClass Gaussian Mixture Modeling (GMM)	Unknown unknowns: – Classification – Special/rare object detection
Outlier Detection or Anomaly Detection	Principal Component Analysis (PCA) K-Means Expectation Maximization Hierarchical Clustering One-Class SVM	Unknown unknowns: – Special/rare object detection
Time-Series Analysis	Artificial Neural Networks (ANN) Support Vector Machines (SVM) Random Forest	Known unknowns: – Novel detection – Trend prediction

Table 2: Applied approaches as well as their applications for the main data mining tasks in astronomy.

5 Data Mining Softwares and Tools

Different scientific areas have similar requirements concerning the ability to handle massive and distributed datasets 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 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.

StatCodes (<http://astrostatistics.psu.edu/statcodes/>) is a web metasite that provides hypertext links to a large number of statistical codes useful for astronomy and related fields. It is being maintained at the Center for Astrostatistics website.

VOStat (<http://astrostatistics.psu.edu:8080/vostat/>) is a simple statistical web-service whose server is located at Penn State University. It is a GUI wrapper in the R language. It not only performs various analyses,

Feature selection/extraction	Applied approaches	Applications in astronomy
Feature Selection	Best First Exhaustive Search Greedy Stepwise Random Search Rank Search Race Search Genetic Search Random Forest ReliefF Fisher Filtering Other wrapper methods	– Reducing dimension – Choose effective features
Feature Extraction	Principal Component Analysis (PCA) Independent Component Analysis (ICA) Linear discriminant analysis (LDA) Latent semantic index (LSI) Singular Value Decomposition (SVD) Multidimensional Scaling (MDS) Partial Least Squares (PLS) Locally Linear Embedding (LLE) ISOMAP Factor analysis Kernel LDA Kernel PCA Kernel Partial Least Squares (KPLS)	– Noise reduction/removal – Reducing dimension

Table 3: Feature selection/extraction methods.

including plotting, data smoothing, spatial analysis, time series analysis, summarization, fitting distribution, regression, many different types of statistical testing, and multivariate techniques, but it also plots interactive 3D graphics. The main goals of VOSTat are to encourage astronomers to use statistics and spread the use of R among astronomers.

Weka (<http://www.cs.waikato.ac.nz/ml/weka/index.html>) implements machine learning algorithms for various data mining tasks, for example, data pre-processing, classification, regression, clustering, association rules, and visualization. Also it can develop new machine learning schemes. It is an open-source, easy-to-use, user-friendly data mining tool useful for data mining tasks from different fields. **AstroWeka** (<http://astroweka.sourceforge.net/>) is a set of extensions to Weka focusing on astronomical data mining tasks. It applies the Astro Runtime and Starlink Tables Interchange Library to load data to and from the Virtual Observatory.

AstroML (<http://github.com/astroML/astroML>) is a Python module developed for machine learning and data mining, which is built on numpy, scipy, scikit-learn, matplotlib, and astropy and is distributed under the 3-clause BSD license. In order to effectively analyze astronomical data, it includes a growing library of statistical and machine learning routines in Python and several uploaded open astronomical datasets and provides a large suite of examples of analyzing and visualizing astronomical datasets. The goal of astroML is to provide a community repository for fast Python implementations of common tools and routines used for statistical data analysis in astronomy and astrophysics and to provide a uniform and easy-to-use interface to freely available astronomical datasets.

DAME (DATA Mining & Exploration) (<http://dame.dsf.unina.it/>) is an innovative, general purpose, web-based, distributed data mining infrastructure, which specializes in massive data sets exploration with machine learning methods. It has been used in astrophysics for: photometric redshift evaluation, photometric quasar candidate extraction, globular cluster search, active galactic nuclei classification, photometric transient classification in multi-band, and the multi-epoch sky survey.

Auton Lab (<http://www.autonlab.org/autonweb/2.html>), directed by Artur Dubrawski and Jeff Schneider, researches new approaches to Statistical Data Mining. It illustrates the directors' great interest in the underlying computer science, mathematics, statistics, and AI of detection and exploitation of patterns in data. There are many talks, tutorials, and software about data mining and machine learning on this website.

Successful data mining toolkits and ideas from other fields or businesses can be borrowed and transformed to astronomy. Philip Bermingham, Data and Analytics Manager, said “We chose Skytree because they’re absolutely leading the way when it comes to Machine Learning”. Skytree Jump Start (<http://www.skytree.net/products-services/jump-start/>) is a tailored approach designed to help users become quickly familiar with machine learning by means of demonstrating/exploring the feasibility of applying machine learning to achieve a specified goal. Skytree deals with large data in linear runtime; therefore it is very useful for big data mining issues. Depending on the latest statistical and computational knowledge, it will help users to develop a machine learning product, service, or capability with game-changing to special goals. Ball (2013) investigated the combination of a cloud computing system, the Canadian Advanced Network for Astronomical Research (CANFAR), with Skytree to create the world’s first cloud computing system for data mining in astronomy. This experiment showed that CANFAR+Skytree has the capability of handling huge data from future large sky surveys such as LSST.

To deal with special astronomical problems, astronomers have developed many toolkits. From the data mining point of view, the photometric redshift measurement of galaxies or quasars is a regression task. Currently, there are many kinds of tools or procedures used for photometric redshift estimation. The Bayesian Photometric Redshift (BPZ) code implements the Bayesian method to estimate photo- z , described by Benítez (2000). Hyperz is a public photometric redshift code based on standard SED fitting procedures, i.e., comparing observed magnitudes with ones expected from using a template of Spectral Energy Distributions (Bolzonella et al., 2000). ANNz is a freely available software package for photo- z estimation using Artificial Neural Networks (Collister & Lahav, 2004). It obtains the relationship between photometry and redshift from an appropriate training set of galaxies with known spectroscopic redshifts. ZEBRA, the Zurich Extragalactic Bayesian Redshift Analyze, combines and extends several of the classical approaches to produce accurate photometric redshifts down to faint magnitudes (Feldmann et al., 2006).

6 Organization

The arrival of the big data era in astronomy has led to a collaboration boom among astronomers, statisticians, computer scientists, data scientists, and information scientists. Faced with difficulties and challenges caused by big data, for scientists, collaboration is the only solution. Because of this situation, various organizations have been established, for instance, the International Astrostatistics Association (IAA, to be affiliated with the International Statistical Institute), the American Astronomical Society Working Group in Astroinformatics and Astrostatistics (AAS/WGAA), the International Astronomical Union Working Group in Astrostatistics and Astroinformatics (IAU/WGAA), the Information and Statistical Sciences Consortium of the planned Large Synoptic Survey Telescope (LSST/ISSC), the American Statistical Association Interest Group in Astrostatistics (ASA/IGA), and the IAA Working Group on Cosmostatistics. These organizations are shown in **Table 4**.

Organization	Under community or project	Foundation Time	Chair
International Astrostatistics Association (IAA)	The International Statistical Institute (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 Ivezić
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 Souza

Table 4: Astrostatistics and astroinformatics organizations.

All the above organizations are described at the Astrostatistics and Astroinformatics Portal (ASAIP) (<http://asaip.psu.edu>), which is a new web site serving the cross-disciplinary communities of astronomers, statisticians, and computer scientists. The ASAIP goal is to support the research of advanced approaches for

astronomy and to popularize such methods into the broader astronomy community. ASAIP provides searchable abstracts to recent papers in the field, several discussion forums, various resources for researchers, brief articles by experts, lists of meetings, and access to various web resources such as on-line courses, books, jobs, and blogs. Researchers and students in astronomy, statistics, computer science, and related fields are welcome to become members of ASAIP. ASAIP is edited by Eric Feigelson (Penn State University) and Joseph Hilbe (Arizona State University). ASAIP is hosted by the Eberly College of Science of the Pennsylvania State University.

7 Conferences and Summer Schools

In order to promote international collaboration and communication in data mining in astronomy, conferences are continually being organized. The Astronomical Data Analysis Software and Systems (ADASS) (<http://www.adass.org/>) conference is held each year at a different hosting astronomical institution. The conference provides a forum for scientists and programmers concerned with algorithms, software, and software systems employed in the acquisition, reduction, analysis, and dissemination of astronomical data. This conference provides a platform for communication between developers and users with a range of expertise in the production and use of software and systems. ADASS XXIV is being held in Canada this year.

Held regularly since 2001, the Astronomical Data Analysis (ADA) (<http://ada7.cosmostat.org/>) conference series focuses on algorithms and information extraction from astrophysical data sets. This conference series has been characterized by a range of innovative themes, including multiscale geometric transforms such as the curvelet transform, compressed sensing, and clustering in cosmology while at the same time it remains closely linked to front-line problems and issues in astrophysics and cosmology. ADA7 was held last year.

The Astroinformatics conference has been held each year since 2010. These four international Astroinformatics conferences have built up a track record of discussion, testing innovative ideas, and building much broader collaborations. These collaborative efforts have formed not just among astronomy institutions but also between astronomy and other disciplines, even including commercial partners, to everyone's benefit. The Astroinformatics 2014 conference (<http://eventos.cmm.uchile.cl/astro2014/>) was held in Chile. Many important new astronomy facilities have been installed in Chile, such as Gemini-South, SOAR (AURA), and VLT Paranal (ESO). Together with the existing facilities, these create an excellent opportunity for the planning and construction of major next-generation optical and radio astronomy facilities, for example, the ALMA Observatory (recently inaugurated), the LSST (Large Synoptic Survey Telescope), and the E-ELT (European Extremely Large Telescope). All these new facilities pose new challenges for massive data flow management. A robust set of interdisciplinary skills are needed to manage and mine tera- and peta-range volumes of data. Astroinformatics 2014 focused on addressing and discussing many of the relevant scientific, technical, and infrastructure issues behind the emerging era of big data in astronomy.

Drs. Jogesh Babu and Eric Feigelson have put forth much effort and done much work on enhancing the dialog between astronomers and statisticians on important research issues. Theirs may be said to be one of the best known cooperation models between statisticians and astronomers. They not only organize many conferences and summer schools but also demonstrate and popularize the application of R language in astronomy (<http://sites.stat.psu.edu/~babu/>). They have been organizing international conferences on "Statistical Challenges in Modern Astronomy" at Penn State every five years since 1991 and an annual summer school there in statistics for astronomers and physicists since 2005. Babu also organized similar summer schools in July 2007, 2008, 2010, and 2013 at the Vainu Bappu Observatory in India. Astrostatistics schools were also organized in 2011 at The Space Telescope Science Institute (STScI) in Baltimore and in 2013 in Valparaiso, Chile.

8 Conclusions

Future advanced facilities will produce unprecedented massive data. Data from different surveys are unique and are needed to do unique science. Integrating distributed datasets from various projects, different times, and different wavelengths will provide large new challenges and opportunities. Some answers to new questions will be obtained by combining disparate sets of information, such as the multi-wavelength characteristics of objects and time-series analysis of variable sources. We live in a big data era, and we should learn about the world with big data views. 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. For storage and computing, cloud storage and cloud computing may be a good solution, but cloud technologies are still in their early stages. For computing,

it is better to move algorithms near data avoiding data transformation because transformation needs a wide internet bandwidth. For management, well-characterized archival data from dedicated surveys and heterogeneous repertoires are of great value, new database technologies are needed, and projects should consider sound data management during the design and implementation period of any facility/instrument. Various databases (e.g., SciDB, MS SQLserver, Oracle, DB2, MonetDB, Vertica, PostgreSQL, MySQL, SQLite, MongoDB, LucidDB, Sphinx, NoSQL) are in bloom, but each has its advantages and disadvantages so care must be taken when choosing what type of database to use. SciDB is a kind of new-born database, piloted by LSST, FermiLab, and other astronomical projects, that increases scientific creativity and easy access to databases and tools as well as the integration and interoperation of tools and datasets that are necessary for astronomers. The speed and efficiency of data mining algorithms directly influence which algorithm to select when performing data exploration. If they are too slow to tolerate when dealing with massive data, the algorithms should be parallelized/distributed by MapReduce or other parallelization technologies on machine learning platforms (e.g., Hadoop, Graphlab, Spark, etc.), indexed by KD-Tree or other index methods, changed into GPU-based or cluster-based ones, or even replaced by other advanced approaches. The big data era has promoted the arrival of an interdisciplinary and multidisciplinary collaboration age. It is impossible to extract the potentially huge value from such massive amounts of data in a reasonable timeframe by depending on single field scientists. If they are not being used at all or are not being used in a timely manner, data are garbage. No one wants to see very costly instruments generating no output. The rise of collaboration among astronomers, statisticians, mathematicians, computer scientists, information scientists, and data scientists is the right solution. Many projects, for example, SDSS, show the feasibility and effectiveness of such collaboration.

Currently, the Sloan Digital Sky Survey (SDSS) (<http://www.sdss.org>) is one of the most ambitious and influential surveys as well as being the most successful and the most cited survey in the history of astronomy. SDSS has evolved into SDSS-I and SDSS-II and has been performing as SDSS-III, working for more than 15 years to make a map of the universe, a task that will continue for many years to come. It has generated more than 5,800 peer-reviewed publications in astronomy and other sciences whose citations have a total of nearly 250,000 hits. Major scientific results of SDSS are the discovery of redshift $z > 6$ QSOs, brown dwarfs and many new gravitational lenses, the sub-structure of the Milky Way, the smallest low surface brightness galaxies, dynamical asteroid families, hyper-velocity stars, and baryon acoustic oscillations. Most events discovered had not even been imagined by the survey developers. SDSS provides unique and interesting data, increasing the amount of precise multi-color photometric data a thousand fold. The data of up to 10^9 objects allow statistical studies to be done in almost any area of astrophysics. The great success of SDSS is due to a number of factors: a large invested research community, data made available as easily as possible, people heavily invested in the program, wide spread exploration while not blindly exploring. The most noteworthy factor is the excellent collaboration among SDSS astronomers, Microsoft, academic computer scientists, and particle physicists, which facilitates fundamental data management and service.

In brief, the success of SDSS gives us the idea that the collaboration among different disciplines and different enterprises represents the general trend in the big data era of astronomy. In the long run, support from governments and communities should be increased and encouraged. At the same time, training next-generation scientists to work with big data is very necessary.

9 Acknowledgements

We are very grateful to referees for their helpful suggestions and comments, which help us improve our work. This paper is funded by 973 Program 2014CB845700 and the National Natural Science Foundation of China under grant Nos. 11178021, 11033001 and NSFC-TAMU Joint Research Program No. 11411120219.


10 References

- Ball, N. M. (2013) *ADASS XXII*, Proceedings of a Conference held at University of Illinois, Champaign, Illinois, USA 4–8 November 2012. San Francisco: Astronomical Society of the Pacific, p 311.
- Ball, N. M. & Brunner, R. J. (2010) *IJMPD* 19, p 1049.
- Benítez, N. (2000) *ApJ* 536, p 571. Retrieved from the World Wide Web November 23, 2014: <http://iopscience.iop.org/0004-637X/536/2/571/fulltext/>
- Bolzonella, M., Miralles, J.-M., & Pelló, R. (2000) *A&A* 363, p 476. Retrieved from the World Wide Web November 19, 2014: <http://webast.ast.obs-mip.fr/hyperz/>
- Borne, K. D. (2009) *Data Mining and Knowledge Discovery Series*, Taylor & Francis: CRC Press, Boca Raton, FL, Ch. 5, pp. 91–114; arXiv/0911.0505.
- Collister, A. A. & Lahav, O. (2004) *PASP* 116, p 345. Retrieved from the World Wide Web November 19, 2014: <http://www.homepages.ucl.ac.uk/~ucapola/annz.html>
- Edwards, K.J. & Gaber, M. M. (Eds.) (2014) *Astronomy and Big Data*, Springer International Publishing: Switzerland.
- Feigelson, E. D. & Babu, G. J. (Eds.) (2012) *Modern Statistical Methods for Astronomy with R Applications*, Cambridge University Press: New York.
- Feldmann, R., Carollo, C. M., Porciani, C., et al. (2006) *MNRAS* 372, p 565.
- Ivezic, Z., Connolly, A.J., VanderPlas, J.T., & Gray, A. (Eds.) (2014) *Statistics, Data Mining, and Machine Learning in Astronomy*, Princeton University Press.
- Li, L., Zhang, Y., Zhao, Y., & Yang, D. (2006) *Progress in Astronomy* 24(4), p 285.
- Murtagh, F. & Heck A. (Eds.) (1987) *Multivariate Data Analysis*, Kluwer Academic Publisher: Dordrecht.
- Starck, J.-L. & Murtagh, F. (Eds.) (2006) *Astronomical Image and Data Analysis*, Springer.
- Starck, J.-L., Murtagh, F., & Fadili, J. (Eds.) (2010) *Sparse Image and Signal Processing, Wavelets, Curvelets, Morphological Diversity*, Cambridge University Press.
- Szalay, A. & Gray, J. (2001) *Science* V, pp 293 & 2037.
- Tagliaferri, R., Longo G., Milano, L., et al. (2003) *Neural Networks* 16, p 297.
- Wall, J. V. & Jenkins, C. R. (Eds.) (2003) *Practical Statistics for Astronomers*, Cambridge University Press: UK.
- Way, M., Scargle, J.D., Ali, K.M., & Srivastava, A.N. (Eds.) (2012) *Advances in Machine Learning and Data Mining for Astronomy, Data Mining and Knowledge Discovery Series*, CRC Press: Boca Raton, FL.
- Zhang, Y., Luo, A., & Zhao, Y. (2004) *Astronomical Data Analysis II. Proc. of SPIE* 5493, p 483.
- Zhang, Y. & Zhao, Y. (2011) *e-Science Technology & Application* 3, pp 13–27.
- Zhang, Y., Zhao, Y., & Cui, C. (2002) *Progress in Astronomy* 20(4), p 312.
- Zhang, Y., Zheng, H., & Zhao, Y. (2008) *Proc. SPIE* 7019, p 701938–1.
- Zheng, H. & Zhang, Y. (2008) *Advances in Space Research* 41(12), p 1960.

How to cite this article: Zhang, Y and Zhao, Y 2015 Astronomy in the Big Data Era. *Data Science Journal*, 14: 11, pp. 1–9, DOI: <http://dx.doi.org/10.5334/dsj-2015-011>

Published: 22 May 2015

Copyright: © 2015 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License (CC-BY 3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/3.0/>.

 *Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS 