# Big Data and Astrophysics

Ricky Carmickle
Indiana University
TBD
TBD, TBD TBD, USA
TBD@iu.edu

## 1 INTRODUCTION

The volume of data generated by astrophysics and astronomical platforms rivals the output of other data sources. Astrophysics and astronomy are considered a primary domain generating 'Big Data', alongside Twitter, YouTube, and Genomics research.[23]

## 2 BIG DATA CHALLENGES IN ASTROPHYSICS

Astronomic data requires perpetual development of data cleaning, storage, processing, searching, mining, and analysis tools.[5]. The data collection tools used for the most data-intensive sky surveys are primarily telescopes, which observe astronomical objects in high definition over a wide range of the electromagnetic spectrum from gamma rays, to visible light, to extremely-low frequency radio waves [21]. The largest astronomic and astrophysics research projects have created databases of hundreds of terabytes, and projects in development are expected to capture data in the exabytes[13, 20, 22].

The Large Synoptic Survey Telescope (LSST), the highest volume astronomical data project currently under construction, is expected to generate 15TB of data per day with over 200 dimensions of data per astronomical object [19]. This 3.2 gigapixel telescope camera is located in Cerro Pachn,Chile and is expected to generate 30 terabytes of data each night of operation for a total of 150 petabytes over the predicted 10-year operational window [12].

The highest-volume astronomical data project currently in the planning stages is the Square Kilometer Array (SKA), which is set to be constructed with portions of the array located in South Africa, Australia, and New Zealand; construction is expected to be completed in 2024. The data collection would consist of multiple radio telescopes in an array design spread over thousands of kilometers. This project would gather 14 exabytes of data each day of operation and store about 1 petabyte of[7] that daily data[11]. Transmission of this data would match the entire data output of the internet in 2013[2].

Making use of this data has challenged almost every part of the Big Data field. In the most basic sense, the processing of data from astronomical and astrophysics platforms is a process of recording high-definition images of the sky, comparing all parts of this image to preceding and successive images to determine the movement of individual objects, then directing the most likely candidates for real astronomical changes to human experts for classification [10]. The depth and detail of images varies depending on the project goals and wavelength of light being observed.

[23]For ground-based observation projects, the most common source of noise in data which requires cleaning are satellites, 'junk' in earth orbit, and defects in the telescope lens which can create artifacts [18]. Two of the most effective methods of cleaning astronomical datasets are the Hough Transform method and the Renewal String Approach [1, 3] [18]. Data curation and storage must be handled in a decentralized way, with researchers and space agencies across the globe contributing to archiving this flow of data to different cloud storage and open source storage systems [9, 16]. The mining of astronomical data has created an entirely new field of 'Astroinformatics'[4] which focuses on efficient management of computing resources. Data mining of astrophysical and astronomical data requires search and selection features which can quickly return data relevant to a researcher's needs[17, 24]. Different query solutions, including SQL, Map-Reduce tools, XtreemFS, and new tools constantly in development[10, 15]. The ability to analyze astronomical data is ultimately a problem of identifying significant events, new attributes for astronomical objects, and interesting "front-page-news" outliers through the "petascale"[6] data containing high levels of noise as well as less significant data.

## 3 HOW BIG DATA HAS CHANGED ASTROPHYSICS

Astronomical and astrophysical data is growing rapidly in size, and researchers are able to gather increasingly large volumes of data as Big Data tool develop alongside the observational technology.

Many of the leading Big Data tools in Astrophysics and Astronomy were developed around the Sloan Digital Sky Survey (SDSS)[7]. SDSS began observations in 1998 and gathered astronomical data as images until 2009. SDSS ultimately gathered 140 terabytes of data which quickly dwarfed the amount of data gathered in the entire history of astronomy. The fields of Astroinformatics and Astrostatistics emerged as data science caught up to this flow of data. The machine learning, data processing, data storage, and data querying were developed concurrent to SDSS and smaller sky surveys like the Palomar Digital Sky Survey, and the SkyMapper Southern Sky Survey [7, 8, 14].

The LSST and SKA are examples of research platforms designed with Big Data tools and methods in mind following the development of these tools with prior surveys. There is little indication that Astrophysics and Astronomy will become less data intensive in the future.

## REFERENCES

[1] et al Amos J. Storkey. 2004. Cleaning sky survey data bases using Hough transform and renewal string approaches. *University of Edinburgh* 347, 1 (Jan. 2004), 36–51.
[2] Ross Andersen. 2012. How Big Data Is Changing Astronomy (Again). (Aug. 2012).

[3] Danko Antolovic. 2008. Review of the Hough Transform Method, With an Implementation of the Fast Hough Variant for Line Detection. *Department of Computer Science, Indiana University, and IBM Corporation* 133 (04 2008), 1539–1548. https://doi.org/10.1541/ieejeiss.133.1539
[4] Kirk Borne. 2009. Scientific Data Mining in Astronomy. *Next Generation of Data Mining (Taylor & Francis: CRC Press)* (2009). https://doi.org/arXiv:0911.0505v1
[5] Kirk Borne. 2014. Top 10 Big Data Challenges fi?! A Serious Look at 10 Big Data Vfis. (2014).
[6] Kirk D. Borne. 2008. A Machine Learning Classification Broker for Petascale Mining of Large-scale Astronomy Sky Survey Databases. *Department of Computational & Data Sciences, George Mason University* (2008).
[7] The Economist. 2010. Special Report — Data, data everywhere. *The Economist* The Economist, February 2010 (02 2010).
[8] G. Jogesh Babu Eric D. Feigelson. 2012. Big data in astronomy. *The Royal Stastical Society* August 2012 (2012), 22–25.
[9] Megan Gannon. 2014. How Scientists Tackle NASA's Big Data Deluge. (01 2014).
[10] et al Harry Enke. 2012. Handling Big Data in Astronomy and Astrophysics: Rich Structured Queries on Replicated Cloud Data with XtreemFS. *Datenbank-Spektrum* 12, 3 (11 2012), 172–181.
[11] IBM. 2012. Square Kilometer Array: Ultimate Big Data Challenge. *SKA Background Information* (2012).
[12] LSST. 2016. Education and Public Outreach (EPO) Completes a Milestone Review, About LSST. (2016). https://www.lsst.org/about
[13] J Tseng R. Newman. 2011. Cloud Computing and the Square Kilometre Array. *Square Kilometer Array* (2011).
[14] et al S. C. Keller. 2007. The SkyMapper Telescope and The Southern Sky Survey. *Publications of the Astronomical Society of Australia CSIRO PUBLISHING* 2007, 24 (2007), 1–12.
[15] M.L. Sellam, T.; Kersten. 2013. Meet Charles, big data query advisor. *UvA-DARE (Digital Academic Repository)* 6th Biennial Conference on Innovative Data Systems Research, CIDR 2013 (January 2013).
[16] Matt Stephens. 2008. Mapping the universe at 30 Terabytes a night: Jeff Kantor, on building and managing a 150 Petabyte database. (10 2008). http://www.theregister.co.uk/2008/10/03/lsst_jeff_kantor/
[17] Matt Stephens. 2010. Petabyte-chomping big sky telescope sucks down baby code: Beyond the MySQL frontier. (11 2010).
[18] Amos J Storkey, Nigel C. Hambly, Christopher K. I. Williams, and Robert G. Mann. 2003. Renewal Strings for Cleaning Astronomical Databases. In *In Uncertainty in Artificial Intelligence 19*. 559–566.
[19] Large Synoptic Survey Telescope. [n. d.]. Large Synoptic Survey Telescope gets Top Ranking, fia Treasure Trove of Discoveryfi. Public Release. (08 [n. d.]). RELEASE LSSTC-09.
[20] Tiffany Trader. 2014. Astrophysics: The Icing on the Big Data Cake. Datanami.com. (01 2014).
[21] Australian National University. 2017. Data Release DR1. (06 2017).
[22] Yongheng Zhao Yanxia Zhang. 2015. Astronomy in the Big Data Era. *Data Science Journal* 14 (2015), 11. https://doi.org/DOI:http://doi.org/10.5334/dsj-2015-011
[23] et al Zachary D. Stephens. 2015. Big Data: Astronomical or Genomical? *PLoS Biology* (2015).
[24] Jacob T. VanderPlas Zeljko Ivezic, Andrew J. Connolly. 2014. *Statistics, Data Mining and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data.* Princeton University Press.

## 4 BIBTEX ISSUES

Warning–no number and no volume in Borne2009

Warning–page numbers missing in both pages and numpages fields in Borne2009

Warning–unrecognized DOI value [arXiv:0911.0505v1]

Warning–no number and no volume in Borne2008

Warning–page numbers missing in both pages and numpages fields in Borne2008

Warning–page numbers missing in both pages and numpages fields in Economist2010

Warning–no number and no volume in IBM2012

Warning–page numbers missing in both pages and numpages fields in IBM2012

Warning–no number and no volume in Newman2011

Warning–page numbers missing in both pages and numpages fields in Newman2011

Warning–page numbers missing in both pages and numpages fields in Sellam2013

Warning–empty publisher in Storkey2003a

Warning–empty address in Storkey2003a

Warning–empty year in LSSTRank

Warning–unrecognized DOI value [DOI: http://doi.org/10.5334/dsj-2015-011]

Warning–no number and no volume in Stephens2015

Warning–page numbers missing in both pages and numpages fields in Stephens2015

Warning–can't use both author and editor fields in Ivezic2014

Warning–empty address in Ivezic2014

(There were 19 warnings)

## 5 ISSUES

DONE:

Example of done item: Once you fix an item, change TODO to DONE

### 5.1 Formatting

Incorrect number of keywords or HID and i523 not included in the keywords

Other formatting issues

### 5.2 Writing Errors

Spelling errors

Are you using *a* and *the* properly?

If you want to say *and* do not use *&* but use the word *and*

### 5.3 Citation Issues and Plagiarism

Remove citations from abstract

### 5.4 Structural Issues

Conclusion missing

Acknowledgement section missing

The paper has less than 2 pages of text, i.e. excluding images, tables and figures