

Rollins College

Rollins Scholarship Online

Honors Program Theses

Spring 2020

Creating a Sample of Off-Color Galaxies Using Big Data Tools

Christopher Becker
cbecker@rollins.edu

Follow this and additional works at: <https://scholarship.rollins.edu/honors>



Part of the [Databases and Information Systems Commons](#), and the [Other Astrophysics and Astronomy Commons](#)

Recommended Citation

Becker, Christopher, "Creating a Sample of Off-Color Galaxies Using Big Data Tools" (2020). *Honors Program Theses*. 130.

<https://scholarship.rollins.edu/honors/130>

This Open Access is brought to you for free and open access by Rollins Scholarship Online. It has been accepted for inclusion in Honors Program Theses by an authorized administrator of Rollins Scholarship Online. For more information, please contact rwalton@rollins.edu.

Creating a Sample of Off-Color Galaxies using Big Data Tools

by

Christopher Becker

A dissertation submitted in partial fulfillment
of the requirements for the
Honors Degree Program
(Interdisciplinary)
at Rollins College
2020

Thesis Committee:

Professor Christopher Fuse, Chair
Professor Daniel Myers
Professor Emily Russell

ACKNOWLEDGEMENTS

A massive thank you to all of the people who helped me throughout this process. Special thanks to each member of my committee for helping me immensely at every stage in this thesis project.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	ii
LIST OF FIGURES	v
ABSTRACT	vi
CHAPTER	
I. Introduction	1
1.1 A Brief Discussion of the Components Necessary for Understanding this Thesis	1
1.2 An Overview of Galaxies	1
1.2.1 Galactic Morphology	2
1.2.2 Off-Color Galaxies	5
1.3 An introduction to Big Data	8
1.4 Combining Astrophysics and Big Data: Astroinformatics.	13
II. Methods	17
2.1 Overview	17
2.2 Determining Parameters	17
2.2.1 Galaxy color	18
2.2.2 Petrosian Ratios	18
2.3 Implementation	19
2.3.1 Initial Implementation Plan	19
2.3.2 Revised implementation	20
III. Results	21
3.1 Results	21
3.1.1 Blue Ellipticals	21
3.1.2 Red Spirals	22

IV. Discussion and Conclusion	24
4.1 Discussion and Conclusion	24
4.1.1 Conclusions from Data on Off-Color Galaxies	24
4.1.2 Conclusions on Astronomy and Big Data	25
V. Future Work	27
5.1 Improving the Experiment	27
5.2 Long-Term work	28
BIBLIOGRAPHY	30

LIST OF FIGURES

Figure

1.1	Hubble Tuning Fork	4
1.2	Hubble-DeVaucouleurs Trident	5
1.3	Hertzsprung-Russell Diagram	7
1.4	Off-Color Galaxy Comparison	8
1.5	Hadoop visualization	13
3.1	Blue Elliptical Query.	21
3.2	Blue Elliptical Results	22
3.3	Red Spiral Query	23
3.4	Red Spiral Results	23

ABSTRACT

Creating a Sample of Off-Color Galaxies Using Big Data Tools

by

Christopher Becker

Chair: Dr. Christopher Fuse

This thesis begins an investigation into the presence of off-colored galaxies in the Sloan Digital Sky Survey. Through establishing the emergence and history of Astroinformatics, the thesis introduces the concepts surrounding both off-color galaxies and the Big Data tools helpful in analyzing the data to find them. A discussion of initial implementation methods and revised implementation due to difficulties with previous plans follows. Results are presented, with well in excess of 500,000 candidates for off-color galaxies present in the sample. Conclusions are then drawn regarding such a large sample and the implications this may have on the conventional understanding of galaxies. Future work and improvements to the project are discussed at length in the closing section.

CHAPTER I

Introduction

1.1 A Brief Discussion of the Components Necessary for Understanding this Thesis

This project requires a functional understanding of several different, but very interrelated subjects. The first aspect centers around an understanding of astronomical concepts, including extra-galactic morphology, on- and off-color galaxies, and the Petrosian ratio. Additionally, an understanding is needed of the Sloan Digital Sky Survey (SDSS) and the data types produced. Finally, a basic level of understanding of Big Data, and the relevant tools employed, such as MapReduce and Hadoop, among others will be established. This introductory section will provide a brief overview of each aspect, while Section 2 will further discuss recent developments and their relevance to the project itself.

1.2 An Overview of Galaxies

This subsection will provide the reader with a working understanding of the properties of Galaxies, including colors, morphologies, and Petrosian ratios. Galaxy morphology is relatively straightforward, with three major classifications, elliptical, spiral, and irregular galaxies *Buta* (1992). For the purposes of this paper, only ellip-

tical and spiral galaxies will be investigated, while irregular galaxies will be filtered out. Elliptical and spiral galaxies are categorized primarily by shape and several elements deemed key to a particular galactic classification, as will be discussed shortly *Doi et al. (1993)*.

1.2.1 Galactic Morphology

In the case of spirals, these galaxies possess a bulge in the center of the galaxy, a disk dominated by young stars, and a halo. Additionally, these galaxies have their namesake spiral arms within this disk. The galactic bulge is usually the oldest part of the spiral galaxy, containing the oldest star clusters in the galaxy *Buta (1992)*; *Graham et al. (2005)*. Typical spirals are identifiable by the dominance of young blue stars in the spiral arms. These stars, of type O and B, are extremely hot, bright, and short-lived, indicating that these stars were produced in recent bursts of star formation *Tojeiro et al. (2013)*

Elliptical galaxies are more easily distinguished due largely to a lack of any defining structure. These galaxies appear as a featureless, vaguely rounded object, often drawing comparisons to a golf ball or egg in shape *Buta (1992)*.

Typically, Spiral galaxies are also identified by a dominance of blue light, indicating star formation and therefore that the galaxy is relatively young. Similarly, elliptical galaxies are typically found to have a dominance of red light, indicating less star formation and an overall older stellar population *Tojeiro et al. (2013)*.

Where the 20th century began with very little understanding of stellar properties and life cycles, the 21st century is poised to serve as a similar, if accelerated timeline for understanding of galactic lifetimes and morphologies. As discussed earlier, the defining features of each class of galaxy have been well-defined, with nearly a century of study to reinforce it. The life cycle and evolutionary timeline of different galaxy morphologies are still being developed, and a focus of much contemporary extra-

galactic research.

When Edwin Hubble first published his findings on galaxies and their varied morphologies, his initial classification system became known colloquially as the Hubble Tuning Fork, seen in figure 1.1. This diagram is the reason for the early- and late-type galaxy monikers *Buta* (1992). In this system, Hubble believed that ellipticals, placed along the "handle" of the tuning fork were the starting point of galaxy evolution. Along the two "prongs" of the Tuning Fork were spiral galaxies that Hubble deemed to be older, more evolved systems. These designations now seem to be backwards, as early-type galaxies (ellipticals) possess older stars, and late-type (spirals) are conversely dominated by recent star formation *Buta* (1992). Hubble's initial belief was that galaxies began as ellipticals and eventually grew into spiral galaxies later in their lifetimes. However, later studies following Hubble's initial publications found that it was physically impossible for the elliptical galaxies to "grow" the characteristic arms of spirals *Buta* (1992). Later work by DeVaucouleurs produced what is called the Hubble-DeVaucouleurs Trident, as DeVaucouleurs added an additional "prong" to the Tuning Fork with the inclusion of irregular galaxies *Buta* (1992). While there has been further work breaking down extra-galactic morphology into more specific descriptors depending on other subtypes of galactic class, much of contemporary discussion on galaxies still utilizes the Hubble-DeVaucouleurs Trident as a basis.

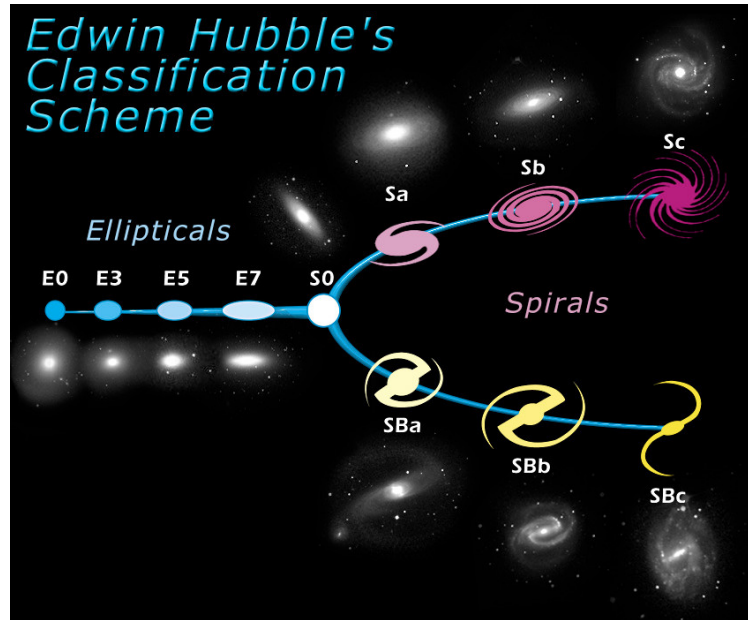


Figure 1.1: The original Hubble Tuning Fork, one of the first attempts to devise a sequence for the evolution of galaxy morphology, the Tuning Fork gained its name from the proposed evolutionary sequences striking resemblance to the two-pronged tool of the same name. Figure courtesy NASA/ESA.

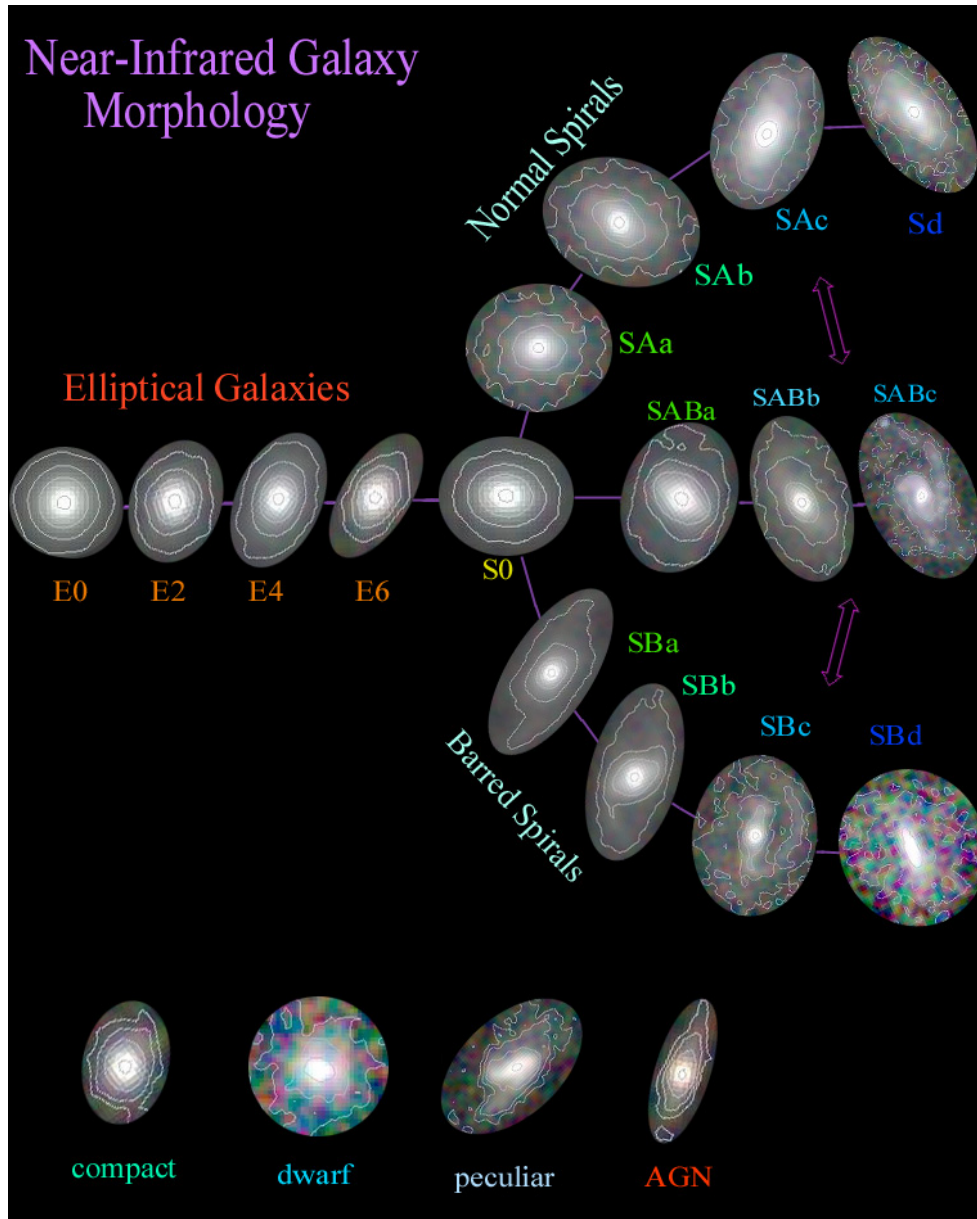


Figure 1.2: The more contemporary Hubble-DeVaucouleurs Trident, which includes the updated classification of galaxies, and subsequently has an additional prong. Figure courtesy NASA.

1.2.2 Off-Color Galaxies

A Recent focus of contemporary extra-galactic research has been in what could be considered outliers, which are galaxies that do not follow known or accepted trends. These atypical cases may help to solidify an understanding of the evolutionary time-

line of galaxies and can help to inspire new hypotheses that may answer questions as to how particular types of galaxies or features of galaxies may come to be. One of the more prominent areas of study as far as outlier galaxies are concerned is off-color galaxies. These off-color galaxies act as outliers to conventional wisdom surrounding the evolution and general ages of different types of galaxies.

Typically, a spiral galaxies are younger than ellipticals, at approximately 1 Gyr to 8 Gyrs in age. as they have had less time to collapse and therefore still have the greater structure. The morphological peculiarities of spiral galaxies, particularly their namesake spiral arms, are prime locations for greater rates of star formation. These spiral arms are abundant with neutral hydrogen, the fuel source from which new stars are formed. Because of this, spirals tend to posses a higher density of young O- and B-type stars which burn much brighter and bluer than many other stats. These very hot stars have significantly shorter life spans than many other stars on the main sequence, in the range of only 500 kyrs *Tojeiro et al. (2013)*; *Schawinski et al. (2014)*; *Bennett et al. (2013)*.

Conversely, ellipticals, which are older, with ages in ranging from 8 Gyrs to 13 Gyrs, having had more time to evolve, have nearly no star-formation. Because of this advanced age, the bright O- and B-type stars so prevalent in spirals have already burnt out and, with the much lower rate of star formation in ellipticals, have not been replaced. This means that the spectra of these older elliptical galaxies are then dominated by the comparatively dimmer and redder, but significantly longer lived K- and M- type stars *Tojeiro et al. (2013)*; *Haines et al. (2015)*; *Bennett et al. (2013)*.

The standard understanding is that spirals are blue and ellipticals are red. Off-color galaxies are then red spirals and blue ellipticals. These off-color galaxies buck the known trend, indicating that the age of these atypically colored galaxies may be different than expected, or that the star formation rates in these types of galaxies has been shifted significantly for some unknown reason. Because of this mystery

surrounding these galaxies, much research has been devoted to understanding what led to these galaxies forming in such odd ways *Tojeiro et al. (2013)*; *Bennett et al. (2013)*.

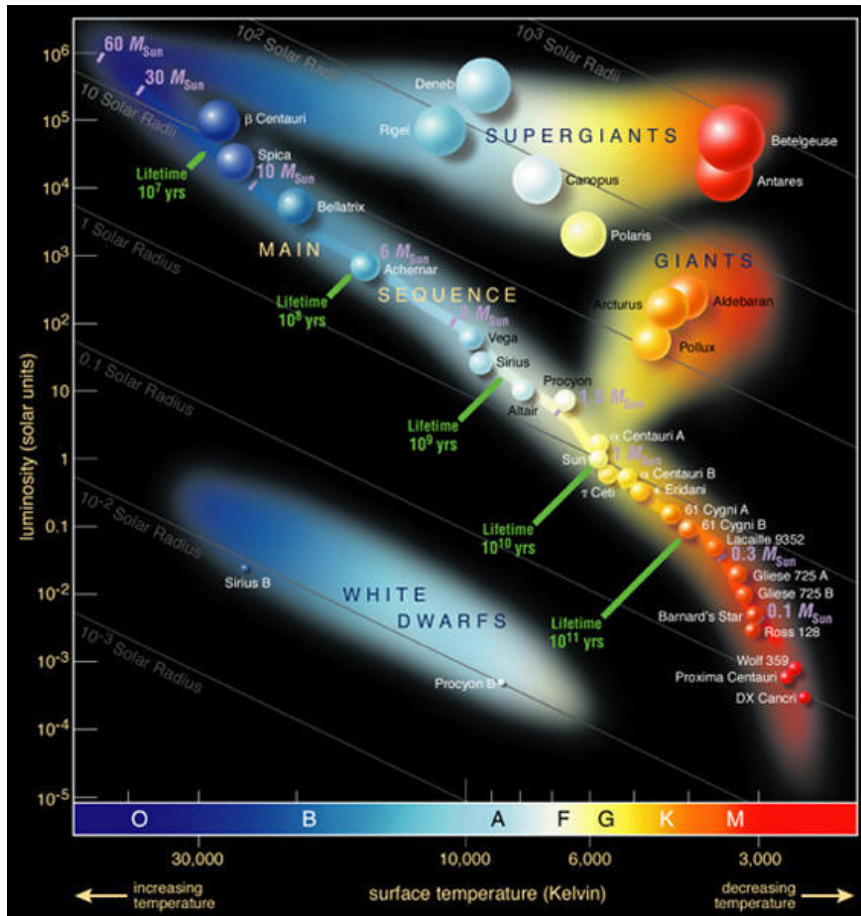


Figure 1.3: Hertzsprung-Russell Diagram (HRD). The diagram shows different stellar classifications along the main sequence, which is the line that connects the hot, bright, young blue stars seen in the upper left to the cold, red, long-lived red stars in the lower right. Standard ellipticals are made up of K- and M-type stars, while blue-light dominated spirals are composed of O- and B-type stars. Image via *Bennett et al. (2013)*

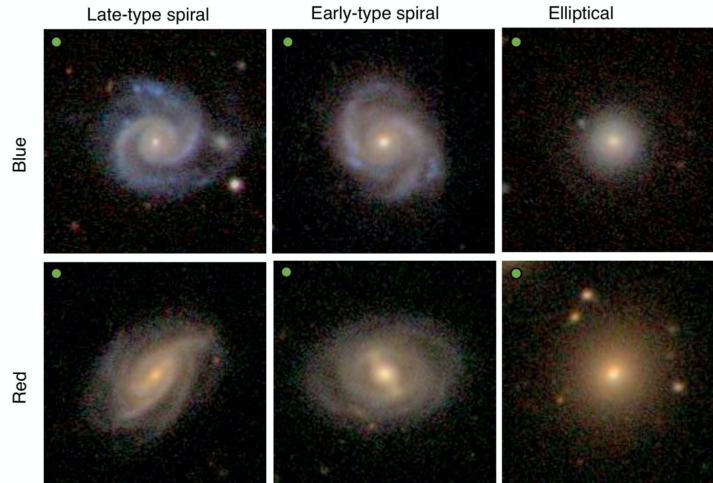


Figure 1.4: A visual comparison of on- and off-color galaxies. Note the on-color blue spirals and red elliptical as well as the off-color red spirals and blue elliptical *Tojeiro et al. (2013)*

1.3 An introduction to Big Data

Big Data is a field of Computer Science that emerged relatively recently, within the past 15 years, and has developed incredibly rapidly as more and more companies and other organizations realize the value of analyzing a variety of data on a massive scale. While the term may seem ambiguous at first, there exist a set of several parameters that qualify something as data fit for analysis by one of several classes of Big Data tools.

The primary qualifiers of Big Data are a set of three aspects known as the three Vs: Volume, Velocity, and Variety *Garofalo et al. (2016)*. Of the three, volume is likely the most straightforward, as it refers simply to the amount of data that exists within a particular database. In current terms, the bare minimum for being a good candidate for analysis by Big Data tools is that the database contains at least on the order of hundreds of Terabytes, where each Terabyte is 1000 Gigabytes. Once more, the volume refers only to the amount of data extant within the system.

The second of the three big Vs is Velocity. Velocity is the rate at which raw data

is generated and added into the database. Velocity also covers the amount of time that it takes to clean or otherwise process the data into a form that is appropriate for use in the database. A simple example of a high velocity data source is Google's search engine, which processes and logs well in excess of 3.5 Billion searches daily, breaking down to 40,000 per second on average, of which each produces a number of different data points including the actual term, top results, time searched, how long the user looked at results, as well as the time the search occurred, among many other aspects *Garofalo et al.* (2016).

The final of the three big Vs is Variety. This is also perhaps the biggest aspect that can qualify a particular database as a prime candidate for analysis by Big Data tools. Variety covers the different types of data within a data base, from sorted to unsorted, to spectra vs query text to even different forms of data i.e. images vs numerical readings vs data cubes. Another major component of data variety is that of the existence of incomplete data sets or even a mix of data types where not every row applies to each individual entry. These types of databases, while oftentimes much more comprehensive, prove especially difficult for analysis in many cases. Because of this, databases with great variety are some of the best fits for analysis by Big Data tools *Garofalo et al.* (2016).

In more contemporary discussions surrounding big data there has been debate over whether to add two additional Vs to the main discussion. These two proposed additions are Value and Veracity. Value is of course referring to the varying levels of value that different pieces of data might provide. A general example being the difference in value between purchase rates of a particular item versus an individual transaction, as there can be significantly greater insight gleaned from a purchase rate versus an individual transaction. The second, perhaps more controversial of the potential additions is the idea of Veracity. Veracity refers to how trustworthy a particular segment of data or dataset is. This includes reliability and accuracy as

well, which may vary depending on how measurement systems or other methods of data collection vary for different categories within the database. Veracity as another Big V is somewhat controversial because veracity may fall under the velocity umbrella since velocity covers time to process and clean data, which may cover the analysis and even filtering of data that may be too unreliable or inaccurate Ora.

An introduction to what might qualify a particular database for analysis by Big Data tools is a great first step, but another key aspect to understanding this whole process is understanding the different classes of Big Data tools. In the context of this paper and its focus on astrophysics, three major categories of Big Data tools will be considered: Machine Learning Algorithms, Data Mining Systems, and the broadest category, Data Analysis tools.

The category of Machine learning algorithms can be further broken down into two subcategories: supervised and unsupervised machine learning algorithms *Baron* (2019). Supervised machine learning algorithms are prepared for full use through a three-stage process. The three stages are training, validation, and testing. The training stage provides the algorithm with a relatively small set of data with well-established target outcomes. This stage helps broadly tune the algorithm to ensure that it produces results in a manner that is consistent and predictable. The next stage, validation, provides the algorithm with a much larger amount of data with known outcomes. Here, fine tuning of the algorithm takes place to help ensure proper sorting by the algorithm and provide further basis that conclusions made from analysis by the algorithm is sound. Finally, the test stage is entered, where machine learning algorithms are fed data sets that are similar to those used to train and validate but without predetermined outcomes. The test stage ensures that the algorithm handles large amounts of data and makes consistent and accurate conclusions from similar data sets. Supervised Machine Learning is perhaps the closest to traditional analysis tools, as they are not too far removed from things such as the solver tool in Excel

which helps to fit a trendline to data. However, supervised machine learning tools are usually far more versatile than something like the Excel solver and can handle significantly higher volumes of data *Baron (2019)*.

An unsupervised machine learning algorithm is a vastly different tool from that of the supervised machine learning algorithm. Where the supervised machine learning algorithms help to find trends in data sets and analyze datasets in a way akin to how researchers would by hand, the unsupervised tools work to find deeper connections between large sets of data that may not be obvious in any amount of traditional analysis. Unsupervised machine learning algorithms find these connections through a varying range of statistical analysis methods, including clustering analysis, dimensionality reduction, visualization, and outlier detection. These algorithms are especially useful in extracting useful data and even conclusions from datasets with very high levels of statistical noise *Baron (2019)*.

The next major category of Big Data tool is Data Mining. These tools perform quite different tasks to those of either type of machine learning algorithm mentioned above. The machine learning algorithms assist in producing models with data or drawing conclusions from data while data mining tools are built more to handle the “cleaning” and processing of raw data. These aren’t built to produce conclusions but to handle large volumes of data and present it in a more readable manner with some work to remove or flag outliers or any data that might seem impossible *Garofalo et al. (2016)*; *Ball and Brunner (2010)*.

Lastly, Data Analytics tools can be discussed. These tools exist as somewhat of a blend of the two previous categories, cleaning up and processing data while also performing rigorous statistical analyses on the database as well. Their wide-ranging application makes them the most versatile and often the most powerful tools one can apply to a database or set of databases. Because of this, they are the tools most often discussed whenever the topic of Big Data comes up.

One of the most prominent tools in the category of Data Analytics is Apache Hadoop, which is an open-source program utilizing MapReduce, a tool built by Google to handle massive-scale data analysis across many networked computers or nodes. MapReduce itself works with a Master computer with a series of worker computers, assigned to either Map or Reduce. Once the master computer has broken the database into manageable chunks, usually on the order of 16 to 64 megabytes, these individual chunks will be sent to a Map worker which will search the chunk for whatever the master computer dictates, and make note of the number of times that particular term or value appears. Once the worker has completed analysis of that chunk, it will send it along to a Reduce node that compiles that worker's report with the reports of many workers before passing its own report along for further condensation into a singular node for analysis *Dean and Ghemawat (2004)*.

Another important aspect of Big Data that is not strictly a Data Analysis tool is the coding language Structured Query Language (SQL). This language is utilized to pull specific sets of data from databases among many other applications in data analysis, both in the field of Big Data and more broadly in many smaller data analysis scenarios. It is built to be versatile and especially in working with databases and database management.

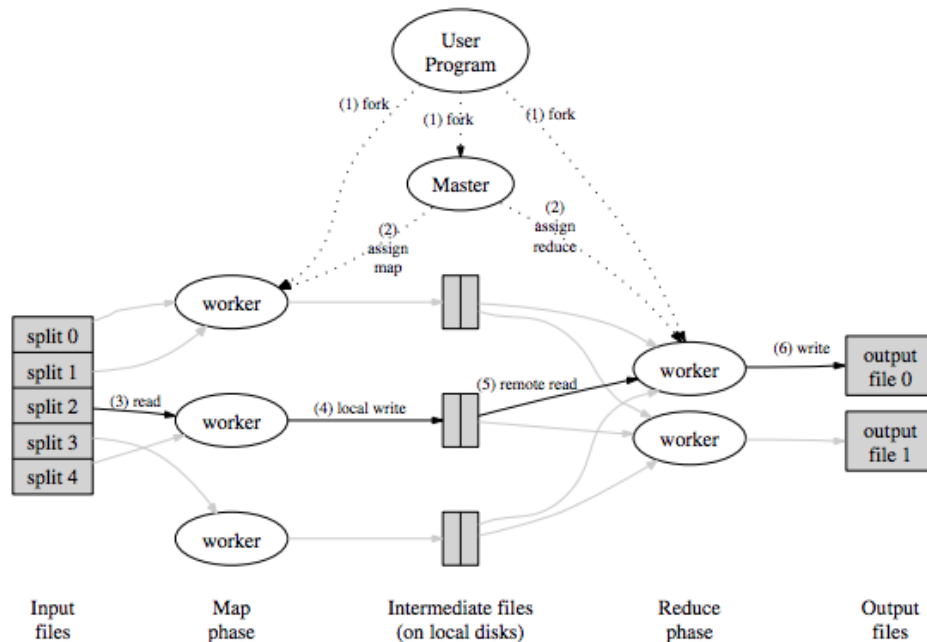


Figure 1.5: A visual representation of the how the Hadoop system functions and distributes the work of the program. Image courtesy Google Research *Dean and Ghemawat (2004)*.

1.4 Combining Astrophysics and Big Data: Astroinformatics.

While these first two subsections have done much to introduce the topics of extragalactic astronomy and big data at individual levels, there is significant overlap in the two fields which only grows as the field of astrophysics continues generating increasingly more data.

One of the best examples of this growing data velocity in the field of astrophysics is to look at one of the more established astrophysical databases, the Sloan Digital Sky Survey (SDSS). A comparison of the data volume and velocity of the SDSS versus predicted volume and velocity of some telescopes and surveys set to come online in the coming decade will be informative. As of Data Release 15, the SDSS contained nearly 200 Terabytes and gained 15 Terabytes between its 14th and 15th releases, which were

approximately a year and a half apart SDS (2018). With no other context, this may initially seem like quite a bit of information on its own and is almost definitely too much for any individual research group, let alone single researcher to comb through in any reasonable amount of time. This is still on the lower end of what might be better suited for analysis by big data programs.

However, the coming decade is projected to see some observatories online with nightly data velocity at or above that of that of the data contained in each SDSS data release *Marshall et al.* (2017). One example is the Vera C. Rubin observatory, also known as the Large Synoptic Survey Telescope, currently under construction in Chile and expected to come online in 2021 and begin scientific observations by the fall of 2022, dependent upon global circumstances. This telescope is projected to produce well in excess of 10 Terabytes nightly, meaning that after just 20 days of observation, the database for the Vera Rubin observatory will be larger than the SDSS in its current iteration, which has been in operation since 2000. The large difference in data velocity can be attributed to the unique style of sky survey that the Vera Rubin observatory will be performing: a synoptic sky survey. Where more traditional surveys like the SDSS will cover significant portions of the night sky, this coverage will occur over several weeks to months or sometimes years. A synoptic sky survey, such as the case of the Vera Rubin observatory will, on the other hand, image upwards of a quarter of the night sky every single night *Marshall et al.* (2017). This has only recently been enabled because of significant advances in sensor technology, data mining tools, and affordable data storage solutions that help to produce workable, easily stored data in a reasonable amount of time without significant issue. These synoptic sky surveys provide a major advantage over more traditional surveys in that they allow observations over time of much larger portions of the sky. Nightly observations of such large sections of the night sky may open up many opportunities for new discoveries and insights into short-term, large-scale behavior of many objects ranging from interior

to the solar system to the far reaches of the universe. The Vera Rubin observatory has 4 main goals: observations relating to Dark Matter and Dark Energy, general observations of rapid events in the universe, studying the structure of the Milky Way and close neighbors, and an inventory of Solar system objects and trajectories. Each of these goals are largely quite different from one another and are on vastly different scales, from within the Solar system to billions of light-years away. This reinforces the idea that synoptic sky surveys are far-reaching and intensely important for new discoveries across nearly every branch of astrophysics.

The data velocity of synoptic sky surveys, such as those produced by the Vera Rubin observatory is not however the largest source of data predicted to come online in the 2020s. Towards the end of the decade, the Square Kilometer Array, or SKA is set to begin observations. This radio observatory, proposed to consist of, unsurprisingly, one square kilometer of networked radio dishes set to observe from either the South African desert or Australian outback, is predicted to have a much greater velocity. At fully operational status, the SKA is set to produce over 150 Terabytes of Data nightly, that is, it will produce three quarters of the amount of data in nearly 20 years of SDSS's operation on a nightly basis. This is a truly vast amount of data compared to any previous astrophysical database and would absolutely require large-scale big data systems to even properly handle and prepare the data nightly. Adding to the complexity of this system is the fact that the SKA, observing in radio and millimeter spectra, which are best suited for 3-dimensional visualization in the form of data cubes, which help researchers identify greater structure in the universe. These cubes are often fed into unsupervised machine learning algorithms to help reduce noise in these cubes and even draw perform analysis and draw conclusions on their own *Hassan et al. (2010); Garofalo et al. (2016)*.

The proposal of these systems coincided with the introduction of a new term and field in astrophysics: Astroinformatics. The international AstroInformatics Associa-

tion (IAIA) was founded and held its first conference in 2010. It exists as a professional organization and touts itself as a bridge field between Data Science and Astrophysics as a whole. In subsequent years, several of the larger astronomical societies, such as the American Astronomical Society (AAS) and the International Astronomical Union (IAU), have also added subcommittees or working groups to themselves to address this emerging field IAI (2020).

This paper contends, however that this is not enough. As discussed above, the data velocity and consequently, data volume, continues and will continue to grow at a blistering pace. There needs to be much more collaboration between big data and astrophysics, and there must be a greater relationship between the fields than a small society and a few subcommittees. Data driven astronomy will be the future and needs to be addressed and moved into the mainstream.

CHAPTER II

Methods

2.1 Overview

The Methods section will be broken into two separate sections. The first section will discuss how the parameters were chosen for determining extra-galactic coloration and morphology, and the second will discuss the initial implementation plans and how the implementation changed throughout the process of collecting the candidate galaxies.

2.2 Determining Parameters

There were several parameters chosen throughout this experiment. The five most important helped to determine the color of the galaxy, the morphology of galaxy, and ensuring that the object in question was in fact a galaxy.

Before the specifics can be discussed, the anatomy of querying the SDSS database and the different terms involved in an SQL query of the database must be outlined. The SDSS is sorted into many different tables, each containing many thousands of columns, with each column pertaining to a particular type of data. For the purposes of this paper, the four major columns of concern for the purposes of finding off-color galaxies are the $petroR50_r$, $petroR90_r$, the Petrosian 50 Percent and 90 Percent radii

in the r-band, respectively, as well as measures of galaxy brightness, $cModelMag_r$ and $cModelMag_g$. The choices will be further elaborated later in the coming subsections.

2.2.1 Galaxy color

Perhaps the most simple of all of the initial parameters was that of the color of the galaxies. This was defined using the g and r band filters from the survey, in accordance with convention and previous work in determining galaxy color from the SDSS. There was further research into which particular g and r band column to choose, which was selected once more based on convention. The final constraints regarding the galaxy colors were determined to be that if the g-r value was greater than 0.6, the galaxy was considered to be red. Similarly, if the g-r value was less than 0.4, the galaxy was categorized as blue. Galaxies lying within the 0.4 - 0.6 range are considered to fall into what is known as the "green valley" and disregarded for the purposes of this experiment.

2.2.2 Petrosian Ratios

The next set of columns to include was that of the Petrosian ratio, which is used in the classification of galaxy morphology. The Petrosian Ratio is based on the Petrosian radius, which is a system for measuring galaxy flux that is built to account for the lack of distinct edges and massively different distances from the observatory to many different galaxies. The Petrosian radius is then essentially a measure of the radius at which a certain percentage of the galaxy's adjusted flux is present. For the purposes of this thesis the 50 Percent and 90 Percent radii are used, with the Petrosian Ratio being r_{50}/r_{90} *SDSS Collaboration and Blanton (2000)*.

With this knowledge in mind, and considering convention set forth by *Fuse et al. (2012)*, the petrosian ratios r_{50}/r_{90} were utilized to enable differentiation between different morphological types. Those galaxies with an r_{50}/r_{90} ratio of less than or

equal to 0.38 were considered elliptical galaxies. Any galaxies with an $r50/90$ of greater than 0.38 were considered spirals.

2.3 Implementation

As alluded to earlier, there was an initial implementation plan that, due to several issues with hardware compatibility as well as with the ways in which SDSS data was made available, had to undergo significant changes. These changes are thoroughly addressed in the revised implementation subsection *Fuse et al. (2012)*; *Graham et al. (2005)*.

While there were some major changes to the implementation plan, the core concepts for classifying off-color galaxies remained the same, based on the parameters outlined in the above subsections. Combining these parameters, the two types of off-color galaxies of interest, blue ellipticals and red spirals, could be identified. This is accomplished by combining the parameters for color and morphology. That is, a blue elliptical would have a $g - r$ value of less than 0.4 and a Petrosian ratio of less than or equal to 0.38 while a red spiral would conversely have a $g-r$ value greater than 0.6 and a Petrosian ratio greater than 0.38 *Fuse et al. (2012)*.

2.3.1 Initial Implementation Plan

The initial implementation plan was to pull all galaxies from the SDSS with a simple SQL query of all objects in the galaxy table, with their corresponding Petrosian radii, g - and r - band magnitudes, as well as object IDs. The actual data analysis and processing would have occurred in Hadoop. This initial plan would have combined the power of Hadoop with the versatility of simple SQL queries to pull all the data in one fell swoop.

In practice, however, a number of limitations and setbacks occurred that prevented the execution of the project in this manner. While each individual limitation would

have likely been easily overcome, the confluence of all of these issues made it easier to move forward with a revised implementation as described in the following subsection. The primary difficulties encountered were as follows. First, across several machines, difficulties were encountered in installation of the prerequisite programs for analysis with Hadoop, stemming from apparent compatibility issues across software versions. Second, and perhaps most impactful, were the limitations placed upon queries to the SDSS - any individual query would result in a maximum of 500,000 individual rows. This was far too small of a sample, as the SDSS contains well over that amount of galaxies, at over 3 Million. Each of these issues combined to produce a situation in which a revised implementation appeared to be most practical.

2.3.2 Revised implementation

The revised implementation was relatively straightforward, forgoing the step of pulling more or less raw data from the SDSS to be analyzed using Hadoop, and instead creating a more complex SQL query that would enable a much more selective way of pulling down data from the database. This revised implementation definitely included application of deeper SQL knowledge in combination with astronomical knowledge to make more informed queries. This was accomplished by entering the parameters discussed above for red spirals and blue ellipticals into the query of the SDSS itself.

CHAPTER III

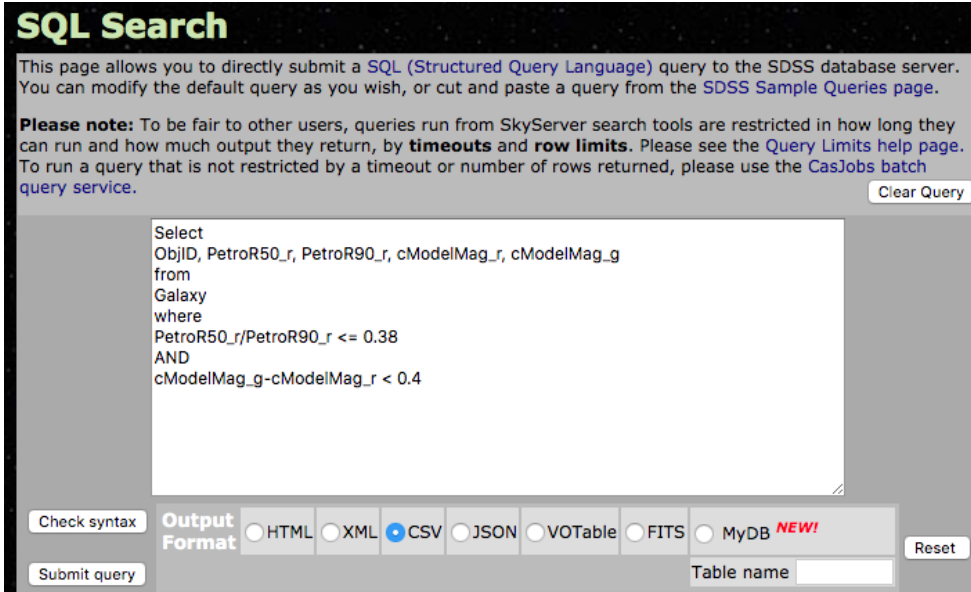
Results

3.1 Results

3.1.1 Blue Ellipticals

The Query submitted to the SDSS in SQL is visible in 3.1.

A small sample of the approximately 500,000 Blue Elliptical candidates this search returned is seen in 3.2.



The screenshot shows the 'SQL Search' interface. At the top, there is a title 'SQL Search' in green. Below it, a paragraph explains that users can submit SQL queries to the SDSS database server and provides links for sample queries and help pages. A 'Please note' section details restrictions on query execution time and output volume, with a link to a 'CasJobs batch query service'. A 'Clear Query' button is located to the right of the note. The main part of the interface is a text area containing the following SQL query:

```
Select
ObjID, PetroR50_r, PetroR90_r, cModelMag_r, cModelMag_g
from
Galaxy
where
PetroR50_r/PetroR90_r <= 0.38
AND
cModelMag_g-cModelMag_r < 0.4
```

Below the text area, there are several controls: a 'Check syntax' button, a 'Submit query' button, and an 'Output Format' section with radio buttons for HTML, XML, CSV (selected), JSON, VOTable, FITS, and MyDB (marked as 'NEW!'). A 'Reset' button is also present. At the bottom right, there is a 'Table name' input field.

Figure 3.1: The query for Blue Ellipticals entered into the SDSS. Note the Selection from the Galaxy table, as well as the inclusion of the restricting parameters.

1	#Table1				
2	ObjID	PetroR50_r	PetroR90_r	cModelMag_r	cModelMag_g
3	1.2377E+18	0.6904899	1.866587	22.29496	22.529
4	1.2377E+18	0.4054346	1.11446	23.38616	23.14851
5	1.2377E+18	0.4841486	1.288301	22.9005	22.93392
6	1.2377E+18	0.9347506	3.743964	23.01745	22.86265
7	1.2377E+18	0.6576223	2.029665	22.54485	22.81079
8	1.2377E+18	0.5851654	1.573983	22.04903	21.44844
9	1.2377E+18	0.8012975	2.319265	22.17124	22.46155
10	1.2377E+18	1.502775	4.448224	21.92226	21.92791
11	1.2377E+18	0.6923824	2.226915	24.48149	24.842
12	1.2377E+18	0.6947798	1.92646	21.89674	20.90925

Figure 3.2: The first 10 rows of the sample are included below.

3.1.2 Red Spirals

The Query submitted to the SDSS in SQL is visible in 3.3.

A small sample of the approximately 400,000 Red Spiral candidates this search returned is seen in figure 3.4.

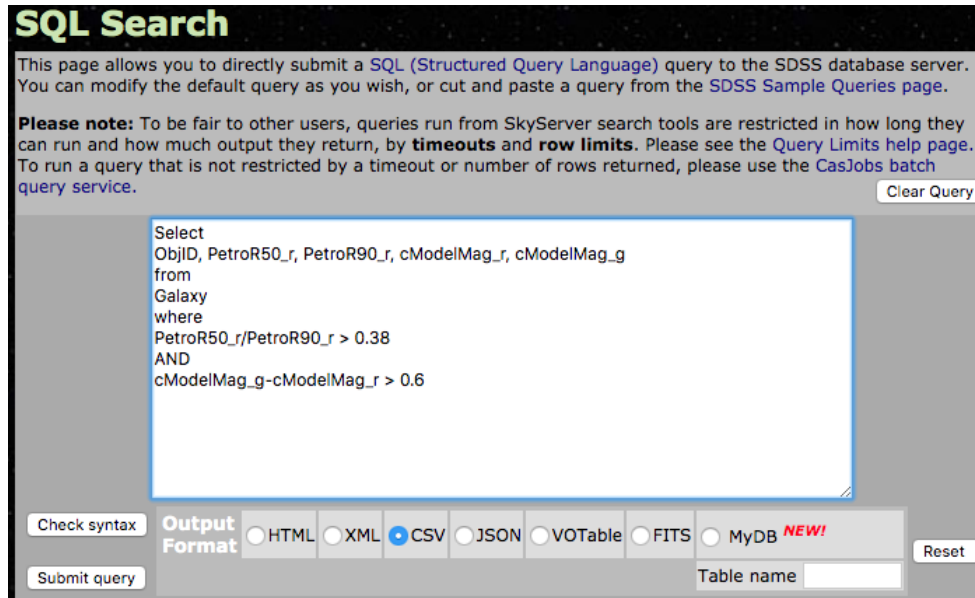


Figure 3.3: The query for Red Spirals entered into the SDSS. Note the Selection from the Galaxy table, as well as the inclusion of the restricting parameters.

#Table1				
ObjID	PetroR50_r	PetroR90_r	cModelMag_r	cModelMag_g
1.2377E+18	1.269057	2.896058	20.02542	20.97435
1.2377E+18	1.828564	3.323782	20.51084	22.20492
1.2377E+18	1.493487	3.253437	20.65777	21.38186
1.2377E+18	2.453389	3.865139	20.34742	21.987
1.2377E+18	1.251989	2.963884	19.61949	20.984
1.2377E+18	1.263184	2.906549	20.39351	21.57381
1.2377E+18	0.8260098	1.680321	21.18296	21.91548
1.2377E+18	0.6211174	0.9352988	21.79368	23.60609
1.2377E+18	0.770783	1.287533	21.63912	23.85526
1.2377E+18	1.469018	2.366769	21.41178	24.72538

Figure 3.4: The first 10 rows of the Red Spiral sample candidates are included below

CHAPTER IV

Discussion and Conclusion

4.1 Discussion and Conclusion

There are two major categories of conclusions to be made from this project. The first is direct conclusions regarding the sample and its implications in the field of astrophysics. These conclusions focus on the importance in the field and the new information gained from this new dataset. The second conclusion will focus on the wider implications of Astroinformatics research and the increasing reliance on Big Data tools the field will face in coming years.

4.1.1 Conclusions from Data on Off-Color Galaxies

Reviewing the results of the project, significant numbers of off-color galaxies are present as compared to what may have been expected. This begs the question of whether or not these so-called galaxies are truly off-color or not. Perhaps, with such a large sample being collected, there exists a wider range of galactic color. Might spiral and elliptical galaxies be morphologically separate from one another, and instead evolving in color? Could this long-standing notion of color association be based solely on observations of the brightest, or oldest and most evolved galaxies? All of these questions can only be answered through further experimentation and analysis of this new sample.

Another important point to be made is that this project appears to be unique in categorizing and finding this many off-color galaxies. Essentially, this is a first-of-its-kind collection of off-color galaxy candidates and provides a very solid base for further investigation into the origins and properties off-color galaxies. This larger-scale collection of off-color galaxies can then allow for a broader analysis of the characteristics of a large population of galaxies, as opposed to the smaller-scale studies of individual galaxies or small samples that a majority of contemporary papers appear to be researching *Tojeiro et al. (2013)*; *Schawinski et al. (2014)*; *Haines et al. (2015)*; *Hao et al. (2019)*.

Furthermore, the large sample sizes fall almost directly in line with what was intended from the outset of the project. This demonstrates well the importance of persisting through setbacks and the advantages of adaptability and flexibility when working on complex projects. Where minimal results may have been achieved using the initial methods laid out, a willingness to adapt and work to find more effective solutions produced favorable results. These results may not have been achieved if the original methods had been followed, given the technical and compatibility limitations that were discovered, as described in the methods section.

4.1.2 Conclusions on Astronomy and Big Data

The level of analysis in the current work would not be possible utilizing more traditional astrophysical methods, due to the sheer amount of data present and the level of analysis needed for each individual galaxy being investigated. The current project, while reasonably novel in its particular focus and application of the Big Data tools utilized, is similar in scope and, in many ways, execution to many possible future experiments (reference the Vera Rubin telescope). This thesis should then act as an example of how to move forward in many projects within the field of extra-galactic research, and even more broadly throughout many different sub-fields of astrophysics.

With the success of this project, and several projects with similar methods, much greater emphasis must be placed on working to increase the presence of Big Data in astrophysics. Promotion of the field of Astroinformatics would seem to be obvious based on the techniques and successes outlined in this thesis.

The expansion of the field of Astroinformatics must occur in order to begin to further modernize the field of astrophysics and allow for more robust analyses of the ever-increasing amounts of data available to researchers. As discussed at length in previous sections, the amount of data available to researchers is already far more than is practical to analyze by more traditional means. With data volumes expected to increase to 150 TB nightly by the end of the decade, as opposed to the contemporary 15 TB in 18 months, the sheer scope of information will extend far beyond impractical and quickly become a necessity to not only use Big Data tools to perform meaningful research on the data, but data mining tools will have to be implemented in order to simply output useful data to a database at the expected data velocity *Garofalo et al.* (2016).

The fast-approaching massive increase in data velocity across nearly all new data sources will bring with it a shift in the role of Big Data in astronomy. As outlined in previous sections and discussion overall, the magnitude of data velocity will necessitate Big Data tools to simply condense the influx of data into something that can be effectively published to a database within a reasonable time-frame. The fact that Big Data tools will be necessary for this first step alone also indicates that more Big Data tools will be required in order to sift through such massive volumes of data. Within the coming decade, Big Data tools will shift from a useful application in performing research on astrophysical data, to a necessity component of processing and analyzing these data. Therefore, a significant shift must occur in the way researchers view and employ astroinformatics techniques.

CHAPTER V

Future Work

There are two major areas for future work to occur regarding the current research and the subsequent results. The first area of future work, is improving the experiment, primarily by focusing on ways to clean the data and better eliminate any false positives or other noise within the dataset. In the middle-term, a better user experience could be developed, which will be discussed in detail. The section immediately following that will then discuss long-term goals for future work. This includes the fact that a wider scope can be taken, building out the program and whatever user experience is developed to allow a wider range of applications within astrophysics, such as some of the other rare examples of purpose-built programs for astroinformatics.

5.1 Improving the Experiment

Perhaps the most significant step to improve this research is the Hadoop analysis used to clean the SDSS data while providing a more in-depth analysis. The major hurdles to implementing a Hadoop improvement were previously discussed. However given more time, the barrier of the 500,000 row limitations of the SDSS query output, as well as other technical and hardware complications and incompatibilities could be worked out. With the benefit of longer time, a more refined data output would be possible using Hadoop's greater analysis and filtering abilities.

One aspect of the Hadoop implementation could include stricter parameters to filter out noise or subprime candidates from being included in the initial dataset. The factor that prevented implementation here, was finding an effective qualifiers for noise and subprime candidates that were easily distinguishable and contained to just one or two parameters within the SDSS. Another step that could be taken would be to filter out high angle galaxies, which may qualify as subprime candidates due to incomplete or improper spectra which lead to a high false-positive rate. This is a specific example of one of the challenges in filtering these poor candidates, as it is difficult to determine effective parameters that are easily accessible solely through SDSS columns.

While the previous two experimental improvements could be implemented in the relative short term, a more medium-term development could be to develop a UI to make analysis easier and more accessible to those unfamiliar with the analysis tools utilized in this experiment. A UI would allow for a broader variety of people to utilize the tools and enact their own restrictions to widen or narrow the search to their liking. The subsequent galaxy samples identified would create a more diverse super-set of samples to allow for more robust cross-referencing ability.

5.2 Long-Term work

In regards to possible longer term steps for this project, a broadened scope has the most potential for researchers. A broader scope could include the development of a UI to perform different and varied analyses beyond the creation of databases of off-color galaxies. There could be expanded work and a shell built around the analysis tools and through the UI to allow greater customization of parameters investigated and restrictions on data and tables. This would potentially allow a much wider range of opportunities for new discoveries, and once more indicate that there is great potential in introducing Big Data tools to a wider range of problems. The success would ensure

astroinformatics would play an ever-growing role in astronomy in general.

That leads into the final step of expanded work: general promotion of the field of astroinformatics. With the success of this project, and if further clean-up and development of the project occurs the project could serve as a signpost for the benefits of astroinformatics and the general increased involvement of data-science in an ever-increasingly data-rich field.

BIBLIOGRAPHY

BIBLIOGRAPHY

- (), What Is Big Data? — Oracle.
- (2018), Data Volume Table — SDSS.
- (2020), About IAIA — IAIA.
- Ball, N. M., and R. J. Brunner (2010), Data mining and machine learning in astronomy, *International Journal of Modern Physics D*, 19(7), 1049–1106, doi:10.1142/S0218271810017160.
- Baron, D. (2019), Machine Learning in Astronomy: a practical overview.
- Bennett, J. O., M. O. Donahue, and N. Schneider (2013), *The cosmic perspective. Stars, galaxies & cosmology*, 7 ed., Pearson.
- Buta, R. (1992), The Morphological Classification of Galaxies, in *Physics of Nearby Galaxies: Nature Or Nurture?*, edited by X. T. Trinh, C. Balkowski, and J. T. V. Trinh, chap. The Morpho, pp. 3–18.
- Dean, J., and S. Ghemawat (2004), MapReduce: Simplified Data Processing on Large Clusters, *Tech. rep.*
- Doi, M., M. Fukugita, and S. Okamura (1993), Morphological Classification of Galaxies Using Simple Photometric Parameters, *Monthly Notices of the Royal Astronomical Society*, 264, 832–838.
- Fuse, C., P. Marcum, and M. Fanelli (2012), Extremely Isolated Early-type Galaxies in the Sloan Digital Sky Survey. I. The Sample, *AJ*, 144(2), 57, doi:10.1088/0004-6256/144/2/57.
- Garofalo, M., A. Botta, and G. Ventre (2016), Astrophysics and Big Data: Challenges, Methods, and Tools, in *Proceedings of the International Astronomical Union*, vol. 12, pp. 345–348, Cambridge University Press, doi:10.1017/S1743921316012813.
- Graham, A. W., S. P. Driver, V. Petrosian, C. J. Conselice, M. A. Bershadsky, S. M. Crawford, and T. Goto (2005), Total Galaxy Magnitudes and Effective Radii from Petrosian Magnitudes and Radii, *The Astronomical Journal*, 130(4), 1535–1544, doi:10.1086/444475.

- Haines, T., D. H. McIntosh, S. F. Sánchez, C. Tremonti, and G. Rudnick (2015), Testing the modern merger hypothesis via the assembly of massive blue elliptical galaxies in the local Universe, *Monthly Notices of the Royal Astronomical Society*, *451*(1), 433–454, doi:10.1093/mnras/stv989.
- Hao, C.-N., Y. Shi, Y. Chen, X. Xia, Q. Gu, R. Guo, X. Yu, and S. Li (2019), Spatially Resolved Studies of Local Massive Red Spiral Galaxies, *The Astrophysical Journal Letters*, *883*, L36, doi:10.3847/2041-8213/ab42e5.
- Hassan, A. H., C. J. Fluke, and D. G. Barnes (2010), Interactive Visualization of the Largest Radioastronomy Cubes, *New Astronomy*, *16*(2), 100–109, doi:10.1016/j.newast.2010.07.009.
- Marshall, P., et al. (2017), Science-Driven Optimization of the LSST Observing Strategy, doi:10.5281/zenodo.842713.
- Schawinski, K., et al. (2014), The green valley is a red herring: Galaxy Zoo reveals two evolutionary pathways towards quenching of star formation in early- and late-type galaxies, *Monthly Notices of the Royal Astronomical Society*, *440*(1), 889–907, doi:10.1093/mnras/stu327.
- SDSS Collaboration, and M. R. Blanton (2000), The Luminosity Function of Galaxies in SDSS Commissioning Data, *The Astronomical Journal*, *121*(5), 2358–2380, doi:10.1086/320405.
- Tojeiro, R., K. L. Masters, J. Richards, W. J. Percival, S. P. Bamford, C. Maraston, R. C. Nichol, R. Skibba, and D. Thomas (2013), The different star formation histories of blue and red spiral and elliptical galaxies, *Monthly Notices of the Royal Astronomical Society*, *432*(1), 359–373, doi:10.1093/mnras/stt484.