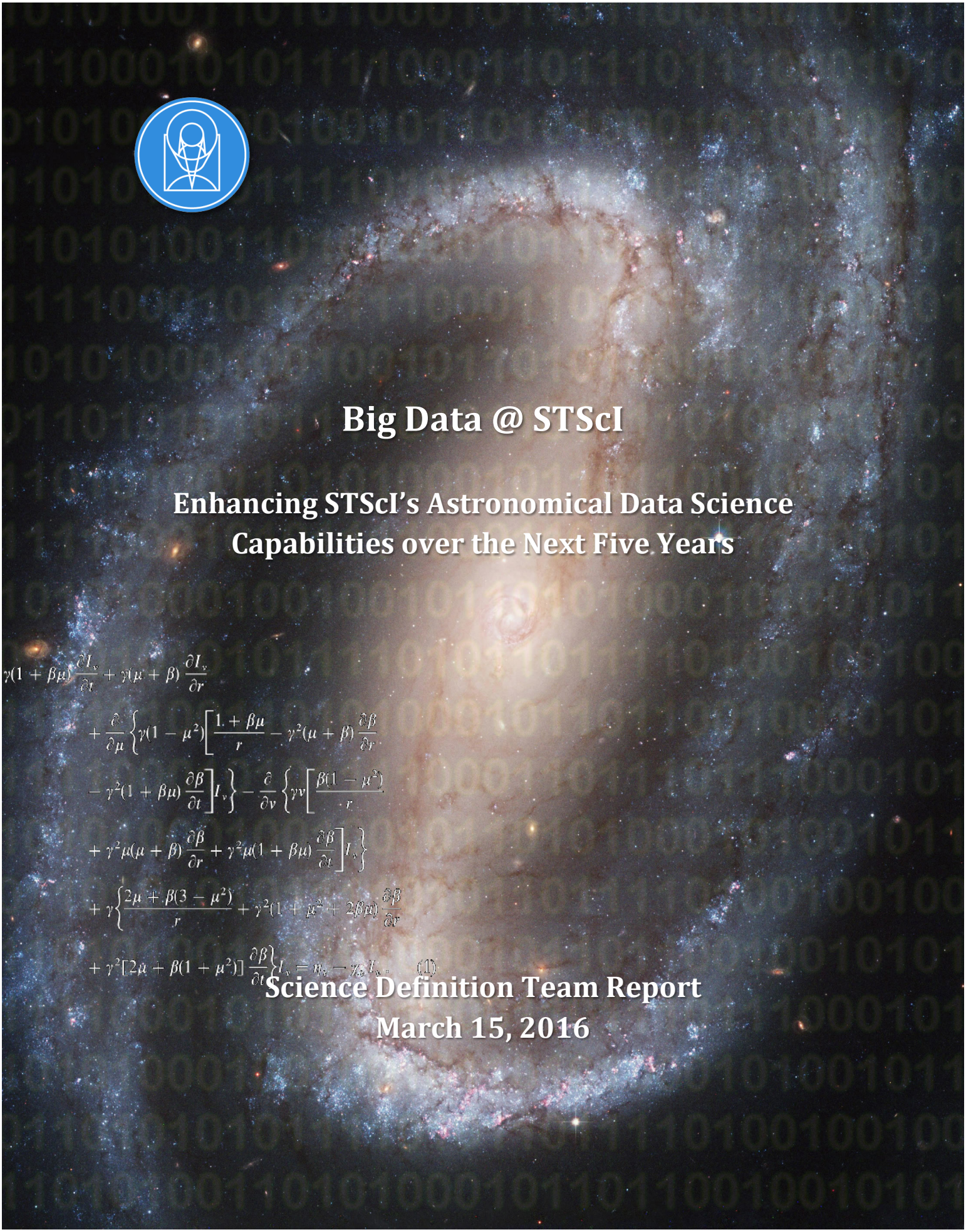

Big Data @ STScI

Enhancing STScI's Astronomical Data Science Capabilities over the Next Five Years

$$\gamma(1+\beta\mu)\frac{\partial I_\nu}{\partial t}+\gamma(\mu+\beta)\frac{\partial I_\nu}{\partial r}$$
$$+\frac{\partial}{\partial\mu}\left\{\gamma(1-\mu^2)\left[\frac{1.+\beta\mu}{r}-\gamma^2(\mu+\beta)\frac{\partial\beta}{\partial r}\right.\right.$$
$$\left.\left.-\gamma^2(1+\beta\mu)\frac{\partial\beta}{\partial t}\right]I_\nu\right\}-\frac{\partial}{\partial\nu}\left\{\gamma\nu\left[\frac{\beta(1-\mu^2)}{r}\right.\right.$$
$$\left.\left.+\gamma^2\mu(\mu+\beta)\frac{\partial\beta}{\partial r}+\gamma^2\mu(1+\beta\mu)\frac{\partial\beta}{\partial t}\right]I_\nu\right\}$$
$$+\gamma\left\{\frac{2\mu+\beta(3-\mu^2)}{r}+\gamma^2(1+\mu^2+2\beta\mu)\frac{\partial\beta}{\partial r}\right.$$
$$\left.+\gamma^2[2\mu+\beta(1+\mu^2)]\frac{\partial\beta}{\partial t}\right\}I_\nu=\eta_\nu-\chi_\nu I_\nu \quad (1)$$

Science Definition Team Report

March 15, 2016

Table of Contents

1 Executive Summary

Astronomical research has entered the era of Big Data. Data spanning many orders of magnitude in wavelength and significantly different spatial resolutions are providing new ways to view the physics of the cosmos. With these riches of data, come the challenges associated with their analyses, either due to the large quantities of data and/or due to the complexity of the data. Taking on these challenges is essential for STScI because our current and future datasets have immense scientific discovery opportunities. It is thus of high strategic importance for STScI to make "data science" a key part of the way we support the science operations of our archive systems and look towards the near future with our best foreknowledge in prioritizing and crafting new and innovative archive user tools and capabilities.

To best assess where and how STScI needs to expand its computing infrastructure, data science skill sets, advanced software environments, and organizational structure, we identified several key science cases that would be difficult, if not impossible, to accomplish with our current capabilities. Based on these use cases, we find that the current scientific computing infrastructure at STScI needs to evolve, and evolve rapidly. Over the next 5 years we will need to grow our data storage capacity by a factor of ~5, increase our internal network bandwidth by a factor of ~10, and provide access to virtualized, dynamically configurable computational power that is ~100 times more capable than our current science clusters.

We also find that a new focus on how we design our archive systems must accompany these hardware enhancements. Forward-looking archive designers must set up enough computing power right on top of the archive and allow users to perform at least the first stages of their astronomical analyses in close proximity to the data. This implies we will need to support and develop server-side application programming interfaces and scripting environments and bring in staff experienced in these methods. We will need to develop and support advanced machine-learning classification algorithms and methods to efficiently reduce the dimensionality of highly complex datasets. And we will require advanced data visualization tools that allow our users to quickly generate scientific hypotheses that can then be more rigorously tested in the science cloud or by downloading the appropriate subset of the dataset.

To further spark new STScI-led innovations in data science, we recommend establishing a postdoctoral fellowship program for early-career researchers who specialize in big data astronomy and the methods needed to extract science from such data. Supporting at least two fellows per year would be ideal. Furthermore, consolidating the leadership of the Institute's archive efforts for all our missions under a single organizational structure will lead to greater coordination, eliminate duplication, enable more efficient use of resources and allow the Institute to take full advantage of synergies between the data science work on our multiple missions.

2 Charter of the Big Data Science Definition Team

In the fall of 2014, the senior management team at the Space Telescope Science Institute (STScI) commissioned an internal Science Definition Team (SDT) to evaluate the scientific opportunities enabled by the availability of very large and/or very complex datasets in astronomy. This action was inspired by several concurrent events: the release of the *Hubble* Source Catalog, the complex data expected to be produced by the *James Webb Space Telescope* (*JWST*) beginning in 2019, NASA's current plans to launch the *Wide-Field Infrared Survey Telescope* (*WFIRST*) in the early 2020s, and by the arrival of the Panoramic Survey Telescope & Rapid Response System (PanSTARRS) science archive data at STScI in 2016.

The SDT was tasked with the following objectives:

- Define at least 5 science use-cases that require extremely large and/or complex datasets and cannot be undertaken today. Emphasis should be given to science use-cases that relate to data that are relevant to STScI's current and potential future missions.
- Discriminate between "big data," "lots of data," and/or "highly complex data" in each use case.
- Identify the functionality of the tools and applications that would be required to accomplish these science use-cases. Do not focus on the implementation or design of such tools.
- Recommend an organizational structure and operational "norms" that optimize efficient and effective implementation for STScI and the user community.

The core Big Data SDT members are ***Alessandra Aloisi, Marco Chiaberge, Gretchen Greene, Anton Koekemoer, Josh Peek, Marc Postman (team lead), Armin Rest, Daryl Swade, Jason Tumlinson, Rick White,*** and ***Brad Whitmore***. In addition, four key advisors to the SDT are ***Karoline Gilbert, Scott Fleming, Joshua Goldberg,*** and ***Dave Liska.***

This report contains the findings of the SDT's 14-month study. Specifically, we identify astronomical datasets and scientific use cases that meet our criteria for "Big Data," which spur us to develop new applications, data structures, services, methods, and architectures for doing the best possible science with the big data of the past, present, and future. We focus on the data science applications and services that enable or that greatly enhance the analyses and dissemination of data currently hosted, or to be hosted in the future, in the multi-mission archive at STScI (MAST). We summarize all our findings and recommendations within this report in its closing chapter.

3 Introduction

3.1 What is Big Data?

"Big Data" is a term borrowed from industry, which collects data on large scale from users and sensors, and is typically used to serve advertisements with a higher return on investment than would otherwise be possible. We will redefine it for this report as data that meets one or more of the following criteria:

- Data whose *raw* form is so large that we must qualitatively change the way in which we reduce, store, and access it.
- Data whose *reduced* form is so large that we must qualitatively change the way in which we interact with and explore it.
- Data whose structure is so complex that our current tools cannot efficiently extract the scientific information we seek.

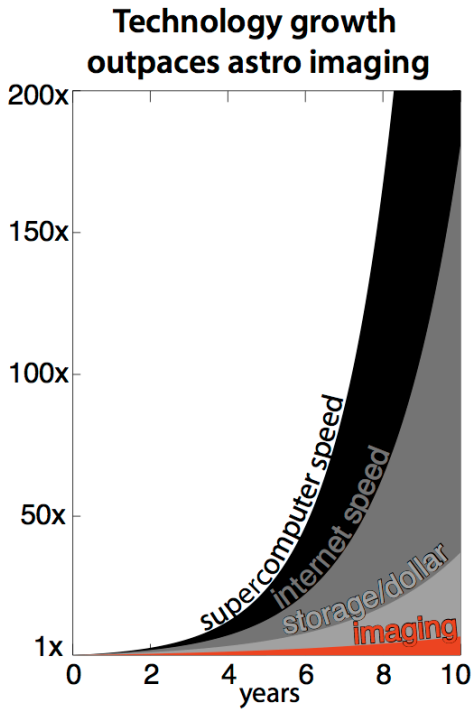

Figure 1: Over 10 years, the rate of UVOIR imaging typically grows by a factor of ~7. In contrast, the amount of storage one can buy, Internet speed, and supercomputer power grow by factors of dozens to hundreds (see https://github.com/barentsen/tech-progress-data).

The technologies that enable the mechanics of data recording, storage, processing, and dissemination grow considerably faster than ultraviolet / optical / near-infrared (UVOIR) astronomical data, as shown in Figure 1. This is to say that, from a technical standpoint, UVOIR big data are actually getting *smaller* in technical cost.

It is not our computers that can't keep up, but our brains: our human ability to explore, comprehend, assimilate, and infer won't keep up with the flow of data unless we build sophisticated tools for interacting with and communicating the ideas hidden in these enormous datasets.

The importance of astronomical Big Data at STScI is its immense scientific discovery opportunities. Throughout this document we will be addressing this issue head on: How can STScI become a leader in extracting *scientific understanding* from the enormous and complex datasets of the future.

3.2 Landscape for Astronomical Big Data: 2015 – 2025

We have entered the Survey Era of Astrophysics. Many different surveys covering large fractions of the sky in all wavelengths either exist (*ROSAT*, *GALEX*, SDSS, PanSTARRS, *IRAS*, 2MASS, *WISE*) or are under way (DES, VISTA, *Gaia*), and their datasets have been used to create comprehensive multi-wavelength catalogs. In addition, bold surveys are being planned for the next decade (LSST, *Euclid*, *WFIRST*, *eRosita*). Some of these surveys include time-domain information (PanSTARRS, DES, LSST, *Kepler*, *TESS, WFIRST, Euclid, Gaia*). Others are complementing multi-band imaging with high-resolution spectra (DESI, SPFS). In addition, some narrow-field observatories (*HST*, *Spitzer*, ALMA, *JWST*, and Extremely Large Telescopes (ELTs) on the ground) have already produced, or will generate, very complex data with high spatial resolution that rivals the richness of a full-sky survey albeit over a much smaller region of sky. The catalogs produced from these surveys are then used for both large-scale statistical studies and for the selection of interesting targets for follow-up as part of guest observing programs. In the era of large multi-wavelength surveys, researchers are likely to ask, "What theories can I test given the data I already have?" at least as often as the more traditional question, "What new data do I have to collect to test a hypothesis?"[1]

A common pattern is that more and more of these large astronomical datasets are placed in easy-to-access archives where individual investigators can perform complex object selections, usually through a query or form-based interface. Access to lower level data (raw, calibrated) is usually also supported. Archive users are getting more sophisticated every year, and their queries (and analyses) are increasingly crossing over archive and wavelength boundaries. By the start of the next decade (2020) this will be even more so, thus the concurrent archive facilities must be built with such considerations and capabilities in mind.

The projected data volume from many of the current and near-future surveys will be a few Petabytes each. Furthermore, the demands of precision cosmology and astrophysics are pushing astronomers to conduct their own analyses of these enormous datasets at the pixel level more than ever before. Hence, efficient methods (in time and cost) for calibrating, accessing, exploring, and analyzing these datasets must be developed in advance. For example, while a single storage system capable of holding a petabyte-scale survey will not be expensive by the end of the decade, transmitting and replicating these data locally at a multitude of user sites will be quite expensive. Cloud computing (as well as dedicated local clouds) should be ubiquitous in the next 5 to 10 years. It is, thus, clear that cloud computing is the direction astrophysics is heading and is a promising approach to conducting science in the survey era of astrophysics. Forward-looking archive designers must set up enough computing power right on top of the archive and allow users to perform at least the first stages of their astronomical analyses in close proximity to the data. This is true for query

[1] Quotes taken from a colloquium given by Zeljko Ivezic at STScI in September 2015.

execution over large catalogs, but will be even more important if users start running their own analysis scripts over the data. Archives of the near future will be more than simple passive file or database servers. Users will rather interact with them algorithmically though well-defined and well-designed Applications Programming Interfaces (APIs). Indeed, such Service Oriented Architectures (SOA) are now becoming the norm for handling complex analytics tasks.

Over the next five years, there will also be key information technology advances relevant to optimizing science-data queries and data analyses. For example, advanced machine learning has undergone enormous advances over the last few years. Robust algorithms, and easy-to-use toolkits are becoming available in common languages, like Python, which makes their reuse extremely easy. Given the complexity of the data fusion demanded by many of the above surveys, we expect that such tools will be widely used by both archive users and the survey pipelines alike. Given the amount of data in the future archives (see **Figure 2** for the specific example of data growth in MAST), we expect that server-side analyses will be commonplace for the users, thus an advanced scripting capability must be supported.

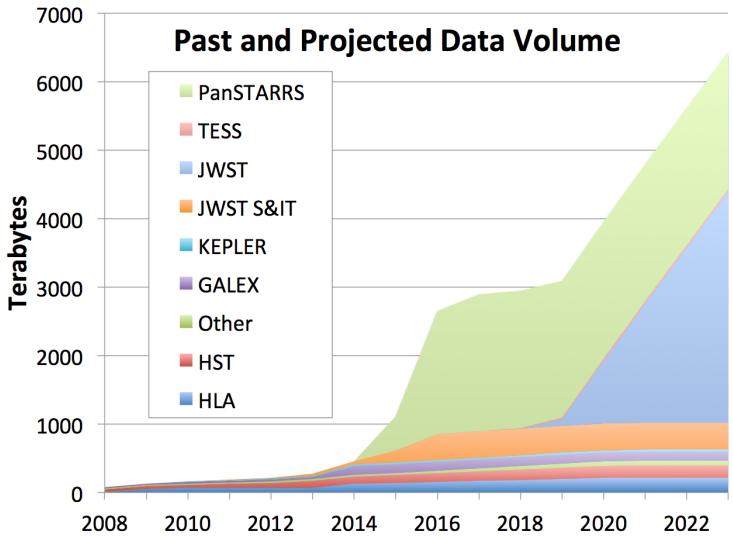

Figure 2: Past and projected growth in the data volume hosted by STScI's Mikulski Archive for Space Telescopes (MAST). In 2016, MAST's holdings exceeded 2.5 Petabytes.

3.3 STScI's Strategic Interests

STScI, as the science operations center for *HST* and *JWST*, has developed a world-class archive with datasets and user interfaces that consistently rank amongst the most widely used in the astronomical community. In addition, STScI is the public science archive center for about twenty other missions, including *IUE*, *GALEX*, and *Kepler*. STScI will also host the public science archive for the PanSTARRS project, whose first data release is expected in 2016. These data are expected to be of comparable popularity to MAST's current holdings, which will change our distribution rate by at least a factor of 4 (see **Figure 3**). In the next 2 to

3 years, STScI will begin storing observations from *TESS* and *JWST*. STScI is also playing a key role in studying how to optimize the science operations of *WFIRST*. Extracting the most science from the observations produced by these missions will often require performing cross-matching and cross-correlating many independent datasets, including data from surveys not currently hosted by STScI. For example, it is already expected that data from LSST, *JWST*, ALMA, *Euclid*, and *WFIRST* will be highly synergistic and will enable science that can only be done by analyzing their observations jointly. STScI thus needs to prepare to not just be a world-class archive but to be a world-class multi-petabyte-scale archive with high performance computing capabilities that support advanced algorithms and data visualization routines to be used by many users. As such, it is of high strategic importance for STScI to make "data science" a key part of the way we support the science operations of our archive systems and look towards the near future with our best foreknowledge in prioritizing and crafting new and innovative archive user tools and capabilities.

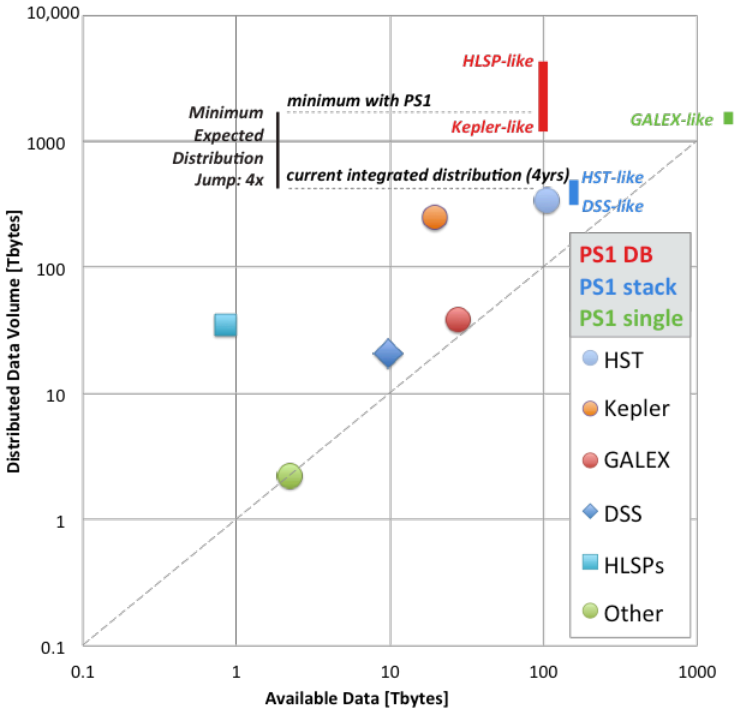

Figure 3: Current and projected data distribution volume as a function of total available data stored. MAST distribution for various data products is shown for the period 2011-2014. Known PanSTARRS holdings for databases, stacked imaging, and single-epoch imaging are shown by the red, blue, and green bars, respectively. The vertical range of the bars represents different guesses for the popularity of the data products, comparing them to the per-bit popularity of current holdings. Even assuming the PS1 data are relatively unpopular as compared to similar holdings, a 4-fold increase in the distribution rate can be expected.

4 Science Cases

To best assess where and how STScI needs to expand its computing infrastructure, data science skill sets, advanced software environments, and organizational structure, we identified several key science cases that would be difficult, if not impossible, to accomplish with our current capabilities. These key science cases are investigations that are central to the scientific objectives of our current or future missions. They can be grouped into a few common challenges in astronomical "Big Data," when applied to large or complex datasets. These are: (1) computer-assisted object detection and classification, (2) multi-wavelength catalog cross-correlations, (3) time-domain event detection and classification, (4) model fitting to massive data volumes, and (5) disentangling complex datasets into their constituent systems. The science cases below provide a representative set of requirements that will ensure that STScI's science archive is able to meet the challenges of data discovery, visualization and exploration in the next 5 years and prepare us for the *WFIRST* era as well.

Over the past six years, major national reviews, including the National Academies 2010 "New Worlds, New Horizons" (NWNH) Report[2], NASA's 2013 "Enduring Quests, Daring Visions" report[3], and NASA's 2014 Science Plan[4], have prioritized the most important astrophysics investigations. These reports all identify three key areas of exploration – the fundamental nature of the universe (e.g., the nature and physics of dark matter, dark energy, baryonic matter), the origin of cosmic structure (galaxies, stars, and planets), and the origin of life in the universe. The NWNH report, in particular, also identifies five "discovery" areas that hold great potential for new knowledge: identification and characterization of exoplanets, time-domain astronomy, high-precision astrometry, the epoch of reionization, and gravitational wave astronomy. The science cases below can all be directly tied back to one or more of these important themes and/or discovery areas.

4.1 Advanced Classification of Astronomical Sources

4.1.1 Classification of Amorphous Sources

An enormous strength of space telescopes is their high angular resolution, unimpeded by the atmosphere. This means more resolution elements per source, leading, generically, to much more spatial information about resolved sources like galaxies and star clusters than ground based photometric data. Encoded in this spatial information is an enormous amount of *astrophysical* information about the sources: their age, their mass, and their star formation histories. Historically, we have attempted to retrieve this information by constructing metrics to reduce this huge image space to a handful of values (feature vectors) like luminosity, light profile shape (e.g., Sersić index), or Gini coefficient (a measure

[2] http://www.nap.edu/download.php?record_id=12951

[3] http://science.nasa.gov/media/medialibrary/2013/12/20/secure-Astrophysics_Roadmap_2013.pdf

[4] http://science.nasa.gov/media/medialibrary/2014/05/02/2014_Science_Plan-0501_tagged.pdf

of non-uniformity in the distribution of light within an astronomical source). We then search for correlations between these features and physical properties measured with spectroscopy and multi-wavelength techniques to make inferences about the structure and evolution of astrophysical objects from the enormous catalogs of images. What follows are a few examples of these experiments, and the opportunity machine vision and machine learning methods present to dramatically expand this work in the era of astronomically big data.

We would like to be able to measure the properties of individual star clusters in nearby galaxies. All stars are born in clusters but dynamical processes disrupt many of these clusters. Studying the population of star clusters in galaxies gives us deep insight into the history of star formation in galaxies and the dynamical forces acting within them. Detecting star cluster candidates in nearby galaxies is not difficult; codes like *SExtractor* (Bertin & Arnouts 2010) or modern Bayesian methods like the *The Tractor* are good at source finding. The difficulty lies in determining the difference between true clusters of stars, and chance overlap of stars in the complex varying background of galaxy disks (see **Figure 4**). At present, the state of the art is for experts to examine and classify these sources by hand, a process that scales poorly as our datasets increase in size. Indeed, the Legacy Extragalactic UV Survey (LEGUS; D. Calzetti, P.I.) team is currently attempting to develop machine learning (ML) algorithms for this application. Their initial outcome – 50% agreement with human classifiers vs. 70% agreement between human classifiers – suggests further optimization of the ML algorithm is needed. But ML methods are the way to go as we progress from classifying thousands of sources to millions of sources.

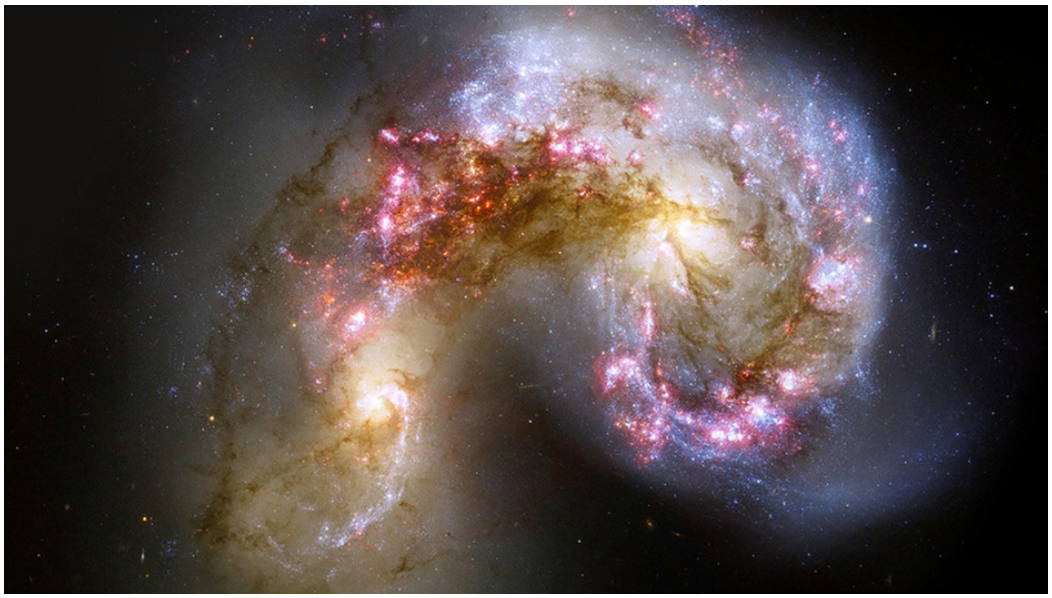

Figure 4: *Hubble Space Telescope* image of two nearby galaxies in the act of colliding (image credit: Whitmore et al. 2010). This galaxy merger triggers star formation and the birth of many young star clusters. Detecting these young star clusters is non-trivial to do in an automated way in the presence of complex and rapidly varying backgrounds.

A similar problem arises on larger scales, trying to classify galaxies by the shape of their stellar halos. Faint stellar halos, stretching out to the virial radius of galaxies, are known to be made up of accretion events over the history of cosmic time, and many of these features remain discernible in deep images of these halos for many billions of years. When a few galaxies are being examined by hand in great detail, it is possible to use modeling techniques to construct accretion scenarios for each of these features, and learn quite a bit about the accretion history of a single galaxy. As high latitude surveys from *Euclid* and *WFIRST* are taken, we will no longer be working on the individual object level, but will want to be able to learn many physical things statistically from much larger samples of halos. In particular, the orbits and masses of accreted galaxies are a sensitive measure of cosmology in the near field. Unfortunately, there exists no way to automate the process of examining galaxy halos and extracting orbital and mass accretion histories from them.

As a demonstration of a science case that can benefit from this, external galaxies are generally well resolved by *HST* since the *HST* point spread function (~0.1") corresponds to physical scales of less than a kiloparsec at all cosmological distances, thus providing ~100x more resolution elements for galaxies ~1" in size than seeing-limited ground-based imaging (where most galaxies above z~0.5 are barely resolved, if at all). In addition, the high sensitivity of typical observations of these sources means that key indicators of morphological disturbances (tidal tails, shells, and other features, signaling the past accretion of smaller systems) are often detected with very good signal-to-noise ratios, and characterizing them more fully would be enabled by these approaches.

Supervised learning has made huge strides towards solving these kinds of problems in the last decade. Supervised learning typically follows a standard script:

1. Gather a subset of your data that have extra information (*e.g.*, spectroscopy) that allow you to label it or directly measure the underlying property of interest. This can also be done through data simulation.
2. Extract from that data a large set of feature vectors (*a.k.a*. columns) that describe the data.
3. Use dimensionality reduction techniques (classically Principle Component Analysis, PCA) to reduce the feature vector space to something both more manageable and meaningful.
4. Feed the reduced vectors to a machine learning classifier or set of algorithms (*e.g.*, random forest – a learning method technique for machine classification that relies on a myriad of decision trees, hence the moniker "forest").

This suite of techniques shows huge promise for solving the problems outlined above: we have large datasets where there is a lot of information in image space, but we do not know precisely where the important and relevant information is. We often have nicely labeled

subsets of data (Step 1) either classified by expert eye, through simulation, or with additional data that allow us to measure the relevant "true" value in a few percent of the images. Step 3 is largely standardized. Step 4, while complex, is well packaged and many different classifiers may be effective.

The stumbling block is Step 2. In the literature, it is not clear what a reasonable set of feature vectors might be for astronomical objects. Often authors use a mix of methods that are partly standard machine vision (wavelets, histograms) methods, and partly methods they suspect could be useful for the astrophysical phenomena under investigation (circles for shells caused by mergers of galaxies, lines for astrophysical outflows from galaxies). But in all the cases above, and indeed in most astronomical applications, the objects follow pretty similar structures. They are centrally concentrated, with a sky that dims towards the edges, are resolved, but often with less than a few hundred resolution elements across. They follow familiar natural shapes, without hard edges or perspective, common to many machine vision feature vector sets.

One way forward would be investment in "learning as a service". This would require maintaining a comprehensive library of feature vectors, connected up to a simplified machine learning apparatus. The service would allow users to supply their own list of object images, either from local repositories or from their own data, and be returned feature vectors, or even a full ML result. We could connect this to online labeling technology for expert classification. Each piece of this stands alone as a useful API for STScI internally and to supply to the community, so that users could use just our feature vector library, or just our learning suite, or both together.

Another possible path forward to efficient classification of amorphous sources in the big data era is investment in Artificial Neural Network (ANN) or "Deep Learning" methods. ANNs can be thought of as very large, stacked, non-linear logistic regressions that are capable of generating very high-quality classifications in absence of feature vectors, i.e. working directly on pixel-level data. Recent methodological innovations have made them the preferred tool for many in the big data industry (Google, Amazon, etc). In particular, Convolutional Neural Networks (CNNs), a concept developed over the last 5 years, work very effectively in classifying images. Dielmann (2015) and Huertas-Company (2015) showed that CNNs can be used very effectively in galaxy morphology identification in SDSS and *HST* data. These networks can be very computationally intensive, and are designed to leverage GPU (graphics processing unit) systems. They have the advantage of being free of feature-vector choice, and thus capable of classifying any kind of astronomical image, but the disadvantage of being more complex in their design. ***We recommend that STScI explore these learning architectures to support the advanced classification and regression problems we will encounter in the big data era***.

The next two science use cases (sections 4.1.2 and 4.1.3) provide examples of challenging classification problems in astronomy that would certainly benefit from advanced "learning services" tuned to datasets in MAST.

4.1.2 Light Echo Identification and Classification

Tycho Brahe's observations of a supernova (SN) in 1572 challenged the teachings of Aristotle that the celestial realm was unchanging. We have discovered a way to see the same light that Tycho saw 440 years ago by observing SN light that only now reaches Earth after bouncing off dust filaments. These *light echoes* (LEs) give us a unique opportunity in astronomy: direct observation of the cause (the explosion) as well as the effect (the expanded remnant) of the same astronomical event. The explosions of low mass and high mass stars are fundamentally different. Many of the heavy elements in the universe are created in these explosive events, and the different types of SNe generate different amounts and varieties of heavy elements. A full understanding of the explosive mechanisms behind the various types of SNe is necessary to trace the evolution of the abundance of heavy elements in the universe.
LEs provide a way to uniquely identify the type of the explosion for historical SNe. Furthermore, multiple LEs from the same supernova allow us to view the explosion from different directions, providing the only way to directly map asymmetries in supernova detonations and shock waves.

The first LEs of ancient events were found serendipitously by Rest et al. (2005) as part of a microlensing survey. They found LEs associated with three 400 to 900 year-old supernova remnants (SNRs). A spectrum of one LE associated with SNR 0509-675 reveals that the SN was caused by the explosion of a low mass star and was a particularly high-luminosity Type Ia SN (Rest et al. 2008a), the first time that an ancient SN was classified based on its LE spectrum. SNR 0509-675 is the first SN for which there are strong constraints on its progenitor (Schaefer, B. et al. 2012), stellar environment (Badenes, C., et al. 2009), the explosion itself, and the eventual SNR (Badenes, C., et al. 2008).

This discovery demonstrated that the LEs of historical SNe and other transients could be found and used to identify the type of star that exploded. Subsequent targeted searches found LEs of Tycho's SN and Cas A (Rest et al. 2008b, Krause et al. 2008a). Follow-up observations of the LEs show that the Cas A and Tycho SNRs were created by a Type IIb SN (the explosion of a massive star) and a normal Type Ia SN, respectively (Krause et al. 2008a, Krause et al. 2008b, Rest et al. 2011a). Analyzing spectra of three Cas A LEs that view the Cas A SN from different directions shows that in one direction the outflow was blueshifted by 4000 km/s relative to the other two LE spectra (see Figure 5). This was the first direct demonstration that the LE phenomenon could allow Earth-bound astronomers to observe a specific individual supernova from dramatically different viewing angles, giving us our first three dimensional view of supernovae remnants. The impact of such information on our understanding of supernovae would be substantial.

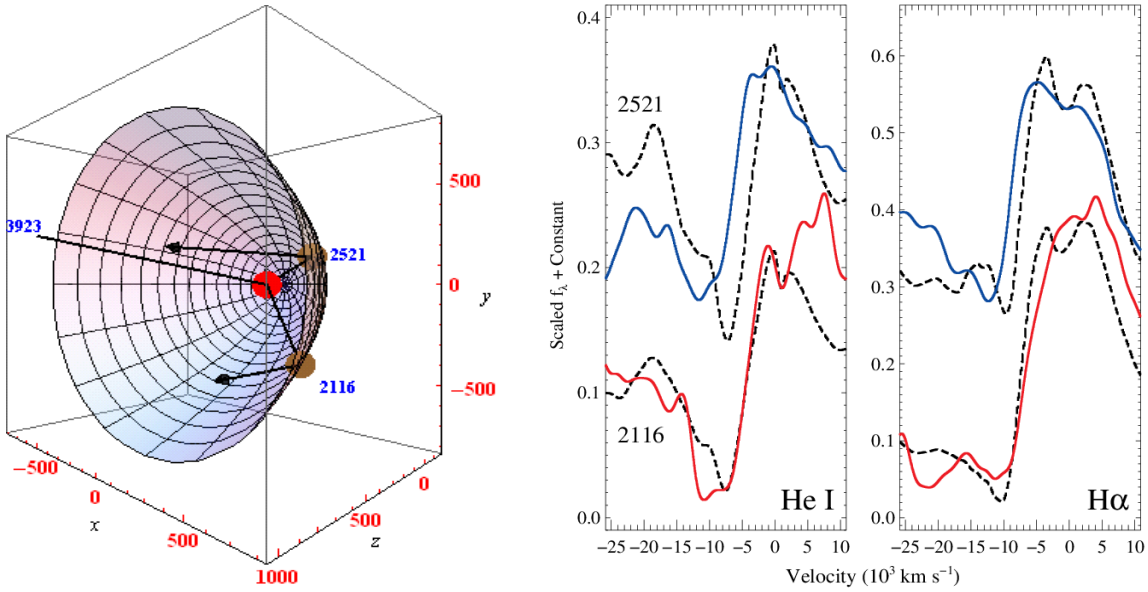

Figure 5: Cas A SN 3D spectroscopy from Rest et al. (2011a, b). Left: Positions of three Cas A SN Les showing how different Les probe the SN from different directions. The SN and the scattering dust are shown as red and brown circles, respectively, in 3D space. (Note: one of the dust locations is not shown for clarity). Right: Line profiles of HeI (587.6 nm) and H-alpha for Cas A Les (colored solid lines) and for light-curve weighted integrated spectra of SN 1993J (dashed black lines).

Current transient searches are providing a flood of extragalactic SNe, and this will accelerate in the era of LSST. These increased numbers provide valuable new information about the statistics of diverse SN types and they reveal exceedingly rare events, but connecting these to the underlying physical parameters is fraught with uncertainty and sometimes impossible. Centuries-old Galactic SN remnants, on the other hand, provide our most detailed and direct measurements of the ejected mass, kinetic energy, element stratification, nucleosynthetic yield, explosion geometry, spatially resolved shock fronts, the surrounding environment, and the nature of any resulting compact remnant, but the original SN explosions themselves were not subjected to the scrutiny of spectroscopic analysis, leaving no evidence of the type of star that exploded. LEs are the only way to bridge this gap. Combining the physical diagnostics of a collection of nearby remnants with LE spectra of the corresponding historical SN event allows us to make solid connections between the underlying physics and observed cosmic explosions.

4.1.2.1 The Big Data Challenge: Reliably Searching for Light Echoes in Large Surveys

Light echoes of historic SNe are, in general, faint (fainter than 22nd mag in visual bands) and at angular distances on the order of degrees to even tens of degrees from the SN position.

Therefore a search area for LEs can span up to several thousand square degrees. The challenge is to cover such a large search area to the required depth. Current and future wide-field time-domain surveys can be used to find light echoes. However, the reference-subtracted single epoch optical images (called difference images) routinely produced by these surveys to identify transient sources do not go deep enough to find the light echoes, and thus seasonal custom stacks and their associated difference images are needed. In addition, conventional point source or galaxy detection algorithms are not suitable for light echo detection. Each of these steps poses its own Big Data challenge.

Table 1: Sky Survey Characteristics

Survey Name	# Filters	Plate scale (arcsec per pixel)	Epochs per year and filter	Total Years	Total # of Tiles	Total # Stacks per filter	Total # Difference images per filter
PanSTARRS	4	0.25	4	4	388,800	1,555,200	1,166,400
PTF	1	1.00	15	4	24,300	97,200	72,900
ASASSN	1	7.80	160	3	399	1198	799
ATLAS	1	1.80	600	3	7,500	22,500	15,000
LSST	4	0.20	20	10	607,500	6,075,000	5,467,500

In order to assess these challenges, we use five current and future surveys as test cases: PanSTARRS (PS1), Palomar Transit Factory (PTF), the All-Sky Automated Survey for Supernovae (ASASSN), the ATLAS survey, and the Large Synoptic Survey Telescope (LSST). The most important characteristics of each of these surveys are described in Table 1. We assume that the images of these surveys are deprojected into tiles with image sizes of 4K x 4K pixel in the native plate scale: deprojections are in general done with plate scales close to the native plate scales in order to conserve the information, but minimize the data storage requirements. The image size of 4K x 4K is also typical; it is large enough to contain enough objects for the difference imaging, and small enough that the processing can be done with over-the-counter hardware. We also assume that these surveys cover 30,000 deg^2. This allows us to estimate the number of tiles, the number of stacks per filter, and the number of difference images per filter. With these numbers, shown in Table 1, we can estimate the data volume, and its requirements on data storage and CPU time.

4.1.2.2 Data Volume, Storage, and Access

An integer 4K x 4K image has a size of 32 Megabytes (MB). An image consists of a flux, variance, and mask frame. Assuming a compression of 0.5 for the flux and variance frame, and a compression of 0.1 for the mask frame, an integer 4K x 4K image has a total compressed size of 35MB. Taking into account the number of tiles, years, and filters, we can

then estimate the disk space needs for each survey (see Table 2). Not surprising, for the surveys with large plate scales (PTF, ASASSN, ATLAS), the disk space needs for the stacks and difference images are small and on the order of a few Terabytes (TB). For the surveys with small plate scales (PS1 and LSST), the stacks and difference images are on the order of 1 Petabyte (PB), which goes beyond simple local storage systems, but is feasible on department or campus-wide computing centers.

Table 2: Disk space and computing time requirements for the different surveys.

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

The input data volume is significantly larger. For PTF and ASASSN, it is still feasible to store the input images locally. However, for ATLAS, PS1, and LSST, the input image set is on the order of a PB (ATLAS, PS1) and 17 PBs (LSST) due to the many epochs and/or small plate scales. In these cases, most likely the input images need to be accessed as needed via the Internet. In particular for LSST, this requires excellent connectivity to ensure the data transfer, data reduction and data analysis can be completed on timescales of just a few months.

4.1.2.3 Data Reduction

In order to reach the depth necessary to find the light echoes, seasonal custom stacks and difference images need to be created. We estimate the CPU time needed to create these 4K x 4K images as follows:

- Stack creation: we estimate that each input image adds 15 seconds CPU time, i.e. a stack created with 10 images uses 150 seconds of CPU time.
- Difference images: we estimate that the difference is created in 200 seconds CPU time.

Table 2 also shows how many CPU days each survey needs for completing the data reduction. ***The data reduction is feasible on a medium sized computing cluster with 100 CPUs for PS1, PTF, ASASSN, and ATLAS. For LSST, a large computing cluster with 1000 or more CPUs is needed***.

There are two options for the location of the data reductions, both with advantages and disadvantages. If the stacking and differencing is done locally, a bandwidth is required that allows the transfer of PBs of data on time scales of weeks or at most months, something that is not currently possible. In addition, significant amount of local disk space and CPUs need to be available as well. The other option is that at least the stacking is done remotely where the original survey data already resides. However, this puts significant strain on the resources of survey, in particular CPUs, disk space, but also on labor. With the exception of LSST, it is unclear whether the other surveys could provide these resources.

4.1.2.4 Object Detection

Detecting rare and faint LEs in a huge collection of multi-epoch, wide-field imaging data is a significant challenge. LEs are faint, extended, and generally arc-like. Their exact shape and size, however, depends on the unique filamentary structure of the reflecting dust, and therefore varies significantly from one echo to the next and with time. The length scales range from a few arcseconds to arcminutes and even degrees. Finding LEs that change from one epoch to the next is therefore much more challenging than identifying transient point sources with PSF-fitting routines, and current codes have not been able to automate the discovery of LEs in a reliable way. Since LEs have an apparent motion of a few arcseconds per year or more, they can be identified painstakingly by eye in difference images. Figure 6 shows examples of LEs from Tycho's SN and Cas A. The brightest LEs have a surface brightness of about 21.5 mag/arcsec^2, but most LEs are significantly fainter with surface brightness ranging from 22.5 to 25.5 mag/arcsec^2.

Another challenge is differentiating real LEs from artifacts of image processing or instrumental effects. Figure 7 shows examples of image artifacts that have similar characteristics to LEs. A typical 4k x 2k image can have several of these artifacts, and therefore over the full survey area of up to 30,000 deg^2 the number of such artifacts can reach hundreds of thousands or even millions. This is in stark contrast to the relatively small number of at most hundreds of LEs we expect to find. For past LE surveys, the data volume was small enough that the LEs could be identified by expert human inspection. However, this dependence on expert humans is not feasible for the current and next generation LE surveys, and artifact rejection has become the bottleneck in the discovery of new LE. The current SN surveys have faced a similar problem, and a significant effort has been invested to solve the problem of discriminating between astrophysical transients and artifacts. For example, PTF successfully uses the technique of recursive feature elimination (Bloom et al 2012, Brink et al. 2013), while a pixel-based random forest classifier was used for PS1 (Wright et al 2015). It is not straightforward to implement these methods for LEs for the following reasons: First, the wide variety in morphology and sizes makes it difficult to define a set of characteristics that differentiates LEs from artifacts. Furthermore, it is computationally much more expensive to compute LE features since the LEs are much larger in size than point sources.

New and efficient methods utilizing a combination of machine learning and citizen science need to be employed to allow the efficient discovery of light echoes in these current and upcoming time-domain surveys. Citizen Science is a method of providing simple repetitive data tasks to the public that are easy for human eyes, but very difficult for machines. Distinguishing between an artifact and a LE is largely trivial to the eye, but it is quite a complex task for a machine. With citizen science application using the newly developed open technology Panoptes on the Zooniverse platform, we can enlist the public in detecting LEs. Users would be shown areas of sky where very simple algorithms seem to have detected echoes in difference imaging. They would then select either `artifact' or `echo' and move on to the next image. These simple, image based tasks have been shown to be very effective and popular in programs like Snapshot Serengeti. This will give us a reliable set of labeled images of echoes. There are, however, millions of these images, too many even for a typical Zooniverse program, since each image should be examined many times for reliable classification.

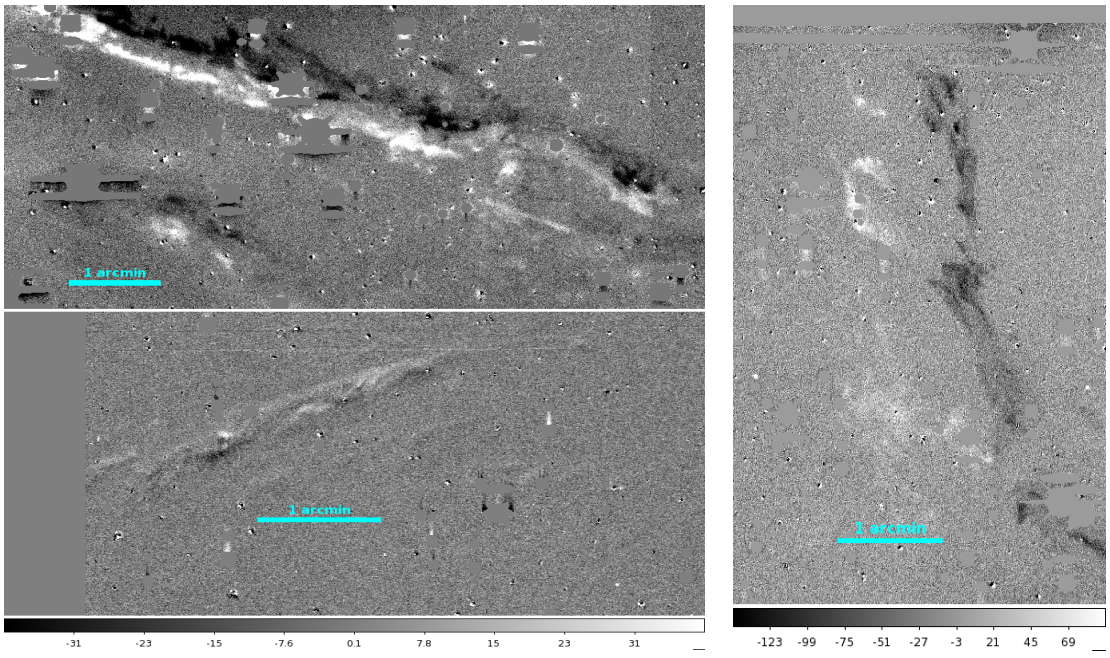

Figure 6: Difference images showing light echoes. The difference image is a subtraction of two images taken at different epochs. The point spread functions of the images are matched prior to subtraction. All static objects, like stars and galaxies, subtract out leaving only excess flux from sources that vary with time. In the above examples the black and white features are changes between the first and second epochs, respectively. Upper left panel: brightest light echo from Tycho's SN. Lower left and right panels: Cas A SN light echoes.

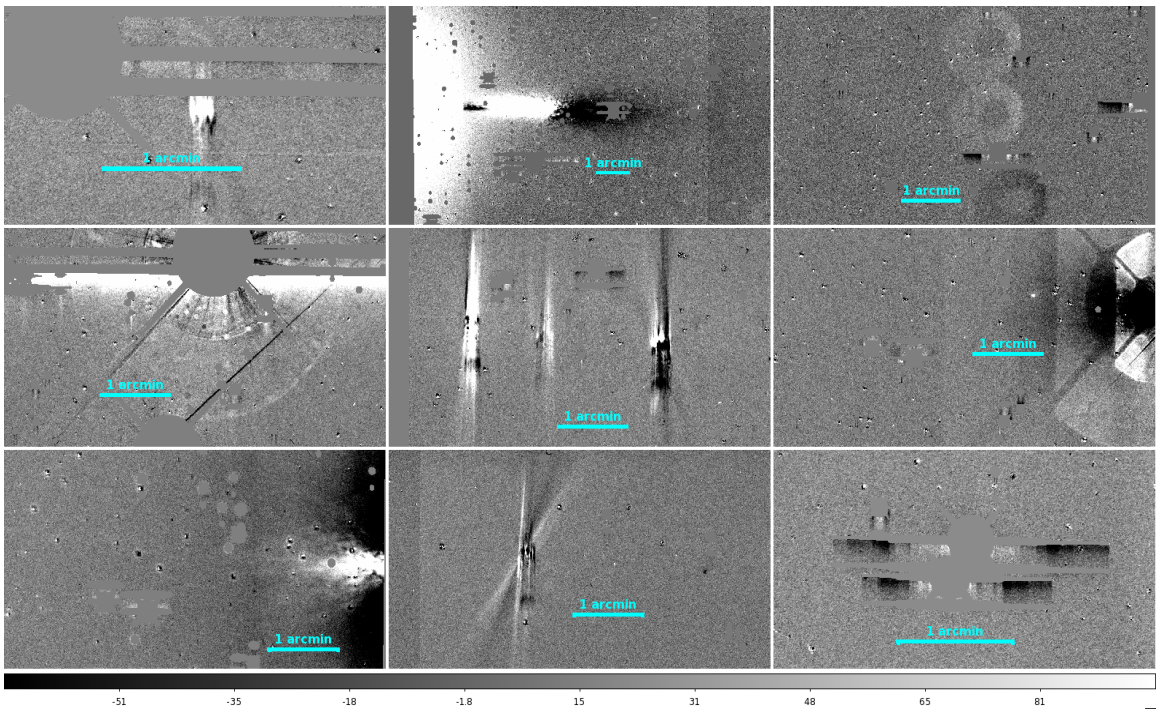

Figure 7: Examples of image and reduction artifacts (stray light, ghosts, diffraction patterns, charge transfer issues) in the difference images, displaying similar characteristics as light echoes.

Machine and Deep learning frameworks have been shown to be extremely powerful for classification of images, distinguishing hand-drawn letters (e.g. MNIST), cats from dogs, etc. The typical procedure is to start with a large number of images with labels; say 10,000 images of dogs, 10,000 of cats. In the case of standard machine learning, feature vectors are extracted from these images. Feature vectors can be thought of various forms of image compression or convolution. Then this space is collapsed using Principle Component Analysis, or similar techniques, to generate lower-dimensional (but information rich) vectors. A subset (say 70%) of these vectors are then fed, with labels, to learning algorithms such as Random Forest, which learn how to classify images based on the labels. Then the classifier can be tested on the withheld 30% of images. In Deep Learning (or Artificial Neural Networks, ANNs) the same process is followed, but the network itself can generate the equivalent of the feature vectors. This has the advantage of being more general, but can be more complex and computationally intensive. Unfortunately, we may not at present have a large enough training set of data to properly train either kind of network. Beaumont et al. (2014) successfully hybridized Citizen Science techniques with Machine Learning to build better classifiers of bubbles in the interstellar medium in the Milky Way. Such an approach will allow us to build up an ANN capable of searching for echoes across the entire area and an efficient citizen science verification network. A workflow diagram for the detection and characterization of SN light echoes is shown in Figure 8.

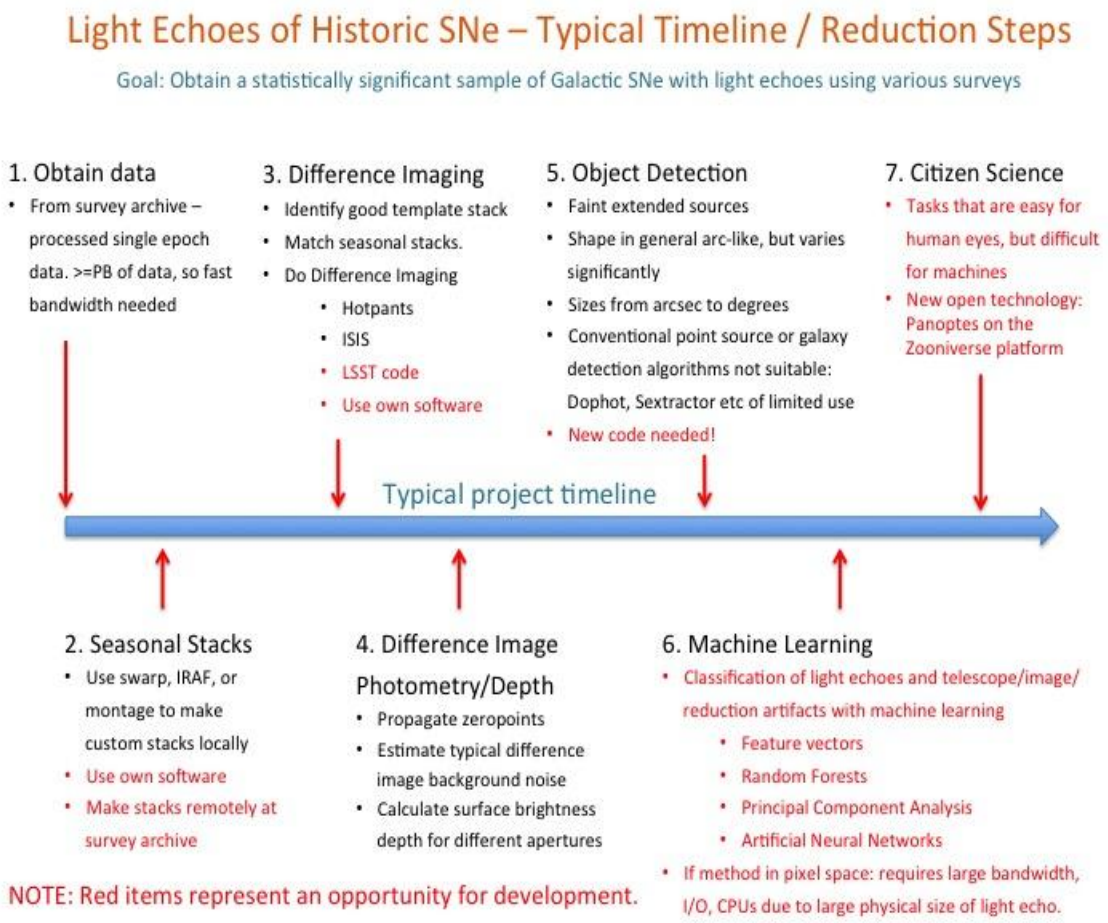

Figure 8: Workflow for the detection and characterization of Light Echoes in astronomical images.

4.1.3 Automated Identification of Gravitationally Lensed Galaxies

Clusters of galaxies are the most massive self-gravitating structures in the universe. They mark the locations of the largest fluctuations in the cosmic mass density field and are predicted to lie at the intersections of large-scale streams of dark matter that make up the cosmic web (see Figure 9). Clusters of galaxies are thus very important structures to study. This is because cosmological models often make substantially different predictions for the number of such rare peaks in the cosmic density field and therefore these models can be tested quite stringently and precisely by measuring the evolution of the cluster mass function (the number density of clusters per unit mass). The two key steps needed to test the models well are (1) identify a large sample of clusters and (2) derive the mass of each cluster in the sample to a reasonable accuracy.

Figure 9: A simulation showing dark matter (blue) and hot gas (orange) centered on a massive cluster of galaxies. The simulation is from the Illustris collaboration (Vogelsberger et al.2014).

The large amount of matter within a cluster of galaxies makes it an impressive gravitational lens that distorts the shapes and positions of more distant galaxies. The degree and extent of the distortions are directly determined by the cluster's gravitational potential. There are two key lensing regions around most clusters – a strong lensing regime (usually within the central 200 kpc of the cluster) and a weak lensing regime (beyond 200 kpc). In the strong lensing regime, gravitational lensing can result in very significant distortion of a

background galaxy's shape and even result in multiple images of the same source at different positions on the sky (see Figure 10). In the weak lensing regime, the shape and position distortions are subtle and can only be detected statistically using large samples of background galaxies. An accurate (±15%) estimate of the cluster's mass can be derived if constraints from both the strong and weak lensing regimes are included in the analysis. Masses derived from gravitational lensing measurements are thus a key component of many current and future cosmological surveys (e.g., Boldrin et al. 2016).

Figure 10: An *HST* image of the cluster MACS J1206.2-0846 obtained during the CLASH program (Postman et al. 2012). The cluster is at a redshift of z=0.44. The *HST* image revealed 47 multiply-imaged lensed galaxies from 12 distinct background sources that span the redshift range 1 ≤ z ≤ 5.5. Even with lensing, many of the magnified galaxies are still very faint.

The *Euclid* and *WFIRST* missions will enable an era of wide-area cluster surveys that provide both deep and high-angular resolution imaging over thousands of square degrees of sky. *Euclid* will survey ~15,000 deg^2 to a depth of about AB mag 24 in the visible and *WFIRST* will survey ~2,300 deg^2 to a depth of AB mag 27 in the near infrared. Estimates suggest that both surveys will detect >20,000 massive clusters of galaxies over a broad redshift range (some estimates put the number as high as 50,000 clusters). The cluster yields from these surveys will enable a precise determination of the redshift evolution of the cluster mass function, providing important constraints on cosmological parameters as well as potentially new constraints on the nature of dark matter itself. The latter is because the mass of a typical cluster is dominated by its dark matter, which can account for up to

80% of a cluster's mass. Because these surveys will have quite good angular resolution (0.3 arcsecond for *Euclid* and 0.11 arcsecond for *WFIRST*) the identification of gravitationally lensed galaxies will be significantly easier than in comparably deep surveys from the ground. But with cluster sample sizes in the tens of thousands, the "by eye" approaches used to identify strongly-lensed galaxies that many researchers have relied upon with small cluster surveys will need to be replaced by automated lensed-galaxy finding algorithms. Lensed galaxies are sometime referred to as "arcs." The number of large arcs in a cluster can be used as a proxy for estimating the cluster's mass (Xu et al. 2015).

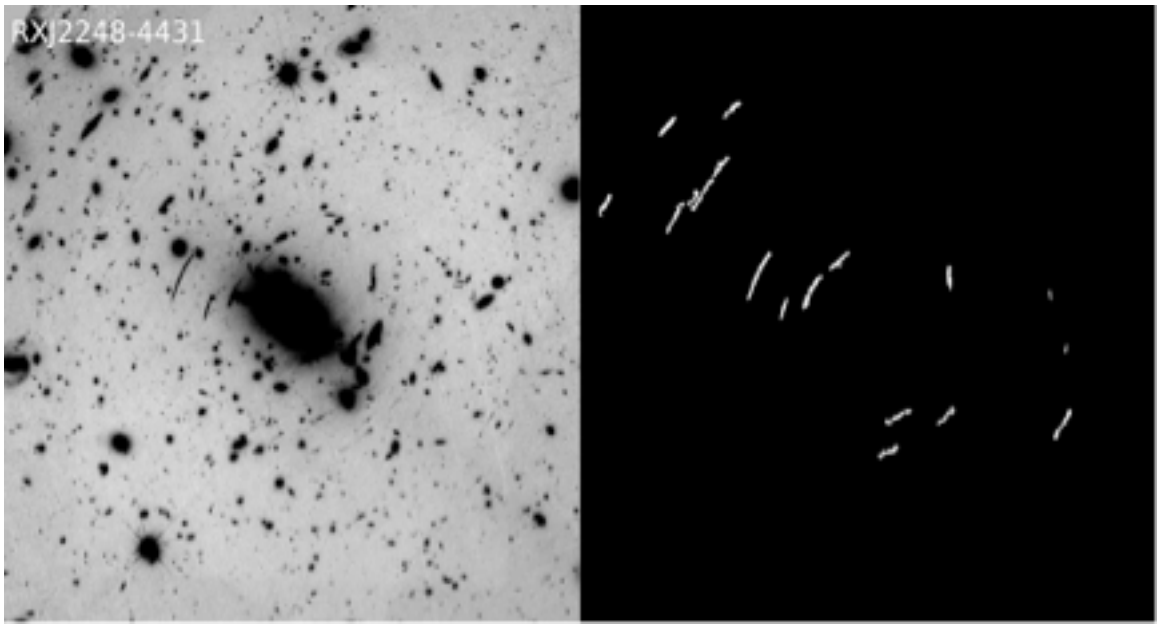

Figure 11: On left: An *HST* image of the galaxy cluster RXJ2248-4431. On right: the identified lensed galaxies in this field as found by the Xu et al. (2015) arc-finding algorithm.

At least two automated lensed-galaxy detection algorithms have been developed to date (Horesh et al. 2005; Xu et al. 2015). The algorithms require access to pixel data as they must perform image filtering and advanced morphology measurements to distinguish likely lensed objects hidden amongst the far more abundant non-lensed objects. An example of the Xu et al. (2015) arc-finding algorithm is shown in Figure 11. This algorithm identifies arcs using their asymmetric intensity gradients as a key discriminator. It is able to detect arcs across a very broad range of brightness with nearly constant efficiency. In the era of *WFIRST* and *Euclid* (and to some extent LSST), we would need a lensed-galaxy finding pipeline that would be able to process the images around tens of thousands of clusters. Based on *HST* results, we would expect to find, on average, about 4 or 5 large arcs (> 6 arc-seconds in size) and at least 10 – 20 smaller arcs in each cluster. In addition, once the arcs are identified, multi-wavelength data would be needed to constrain their redshift as the arc's position and the arc's distance from the cluster are needed to derive a robust mass model. Finally, the arc detection efficiency and false positive rate would need to be

measured using the observational parameters for each survey. This latter step is essential in order to compute the true intrinsic arc abundance rate from the raw detection rate. A workflow diagram for arc detection in a large area survey is shown in Figure 12.

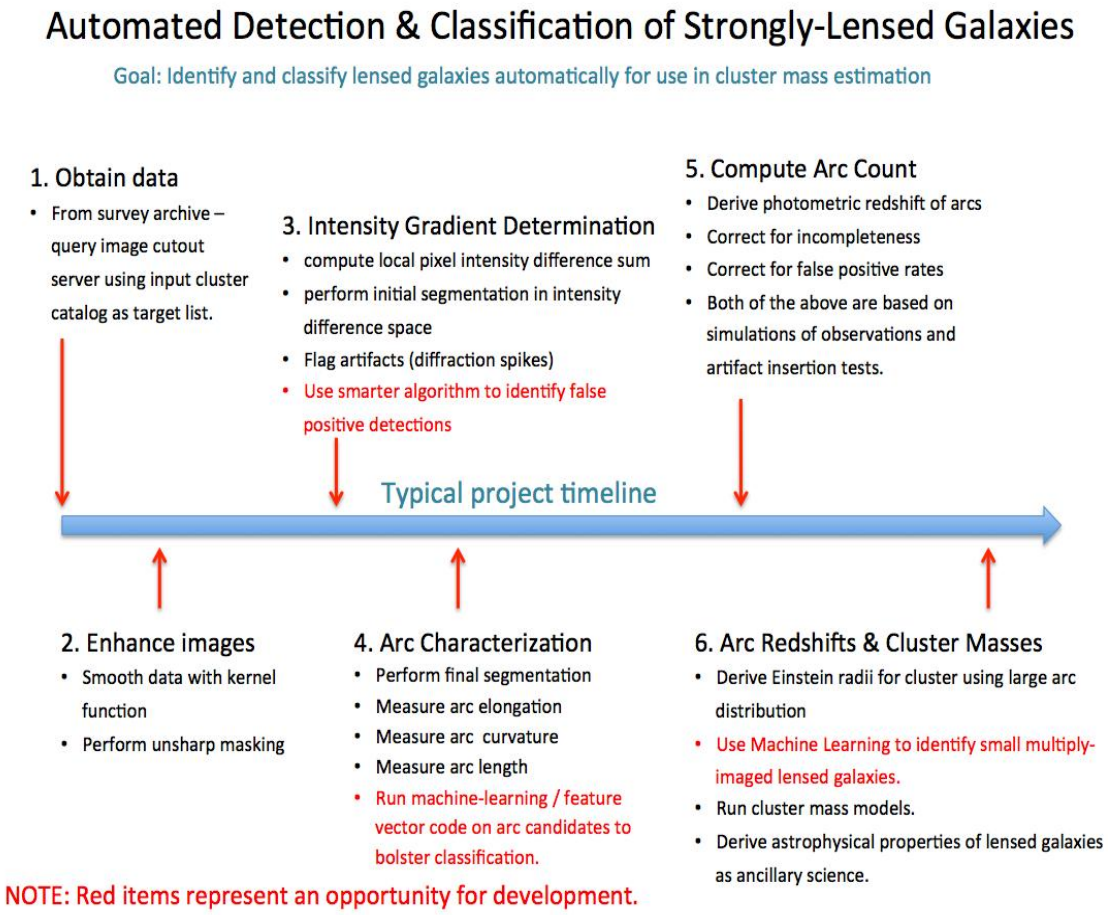

Figure 12: Workflow for the detection and characterization of strongly lensed galaxies behind massive galaxy clusters. Advanced machine learning algorithms combined with high-performance computing systems would enable an analysis of the mass distribution of tens of thousands of clusters expected in all-sky space-based surveys.

4.1.3.1 The Big Data Challenge: Lensed Galaxy Classification

The main computing driver for implementing automated arc-finding on large high-angular resolution sky surveys is access to a high-performance (many cores) computing system. Current arc-finding algorithms require up ~300 CPU-seconds (on current technology desktop workstations) to detect and characterize the arcs in a 3000 x 3000 pixel image. Upcoming space-based surveys like *Euclid* and *WFIRST* are expected to detect between 20,000 to 50,000 clusters, over a range of redshifts. While not all will be strongly lensing clusters, one will still want to run arc-finding algorithms on all of them to determine this in an objective manner. To do this would thus require ~1,700 to ~5,000 CPU hours (2 to 7 CPU months). However, the arc-finding process is highly parallelizable as each

cluster can be analyzed independently. Hence, a cluster with 1000 cores could reduce the clock time of executing the full survey arc-finding analysis to just a few hours. This would allow one to experiment with different detection criteria and also to run the algorithms on the corresponding simulated datasets (needed to establish the completeness and false positive rate). These algorithms may also be optimized for use with GPU-based computing infrastructure but that may not be absolutely necessary. The underlying assumption here is that the pixel data for the central regions of all clusters is available. This would require only about 2 to 3 Terabytes of storage - and is, thus, not considered to be a major technology driver.

In sum, the main computing infrastructure augmentation to support full-sky cluster arc-finding algorithms would be access to all pixel data in the central regions of clusters and the availability of a ~1000 core processor array.

4.2 Resolved Stellar Populations

Observations of various populations of stars are essential for a wide range of key astronomical studies. On large scales, an understanding of stellar populations is necessary for interpreting measurements of the integrated light of distant galaxies, which are too far away to allow us to study their individual stars, in order to learn how galaxies evolve over time. On intermediate scales, stellar populations enable the study of star clusters and stellar associations, which are the environments where most of the stars are actually born. On small scales, detailed studies of individual types of stars allow us to test our understanding of how stars form and evolve. The key observations in all these cases are simple photometric measurements; how bright is the light in a number of photometric bands. For distant galaxies, only integrated light from large numbers of stars can be studied. For nearer galaxies, as discussed in this section, the light from individual stars is measured to study their "resolved"[5] stellar populations.

To resolve individual stars with the highest degree of photometric accuracy often requires observations from the *Hubble Space Telescope* (and, in the near future, with *JWST, WFIRST,* and large ground-based telescopes with advanced adaptive optics). The most common diagnostic tool is the color-magnitude, or Hertzprung-Russell (HR) diagram, an example of which is shown in Figure 13 for a galactic globular cluster. Until about a dozen years ago, astronomers thought they had a pretty good idea of how stars in globular clusters formed and evolved. Globular clusters were believed to be simple stellar populations (SSPs), all formed at the same time with the same chemical composition. However, recent multi-population HR diagrams of many globular clusters, such as Omega Cen in Figure 13, show the presence of several populations. This has been a major surprise and shows that we still have a lot to learn!

Creating the data needed to construct a color-magnitude diagram for a star cluster can be challenging and the study of resolved stellar populations is often computationally intensive. Images of globular clusters typically have hundreds of thousands of stars. The fields are often very crowded, requiring the use of iteratively determined PSF-fitting photometry to optimize the measurements. In addition, a large number of repeat measurements are commonly observed to observe a wide dynamic range (short exposures for bright stars and long exposures for faint stars), variability of the stars, and proper motions. For example, there are over 2000 observations of Omega Cen in the Hubble archives.

Observations of resolved stellar populations in nearby galaxies can be used to determine the detailed history of star formation in a galaxy. An example is shown in Figure 14. This is sometimes called "near-field cosmology" as it allows us to constrain events that occurred in the early epochs of the universe using detailed observations of stars in nearby galaxies. For

[5] Here the term “resolved” means individual stars are detected in an image. The stars themselves are, however, spatially “unresolved” objects, even with our largest telescopes.

example, the Williams et al. (2015) study highlighted in Figure 14 reveals evidence for a global burst of star formation in the Andromeda galaxy (M31) two to four billion years ago.

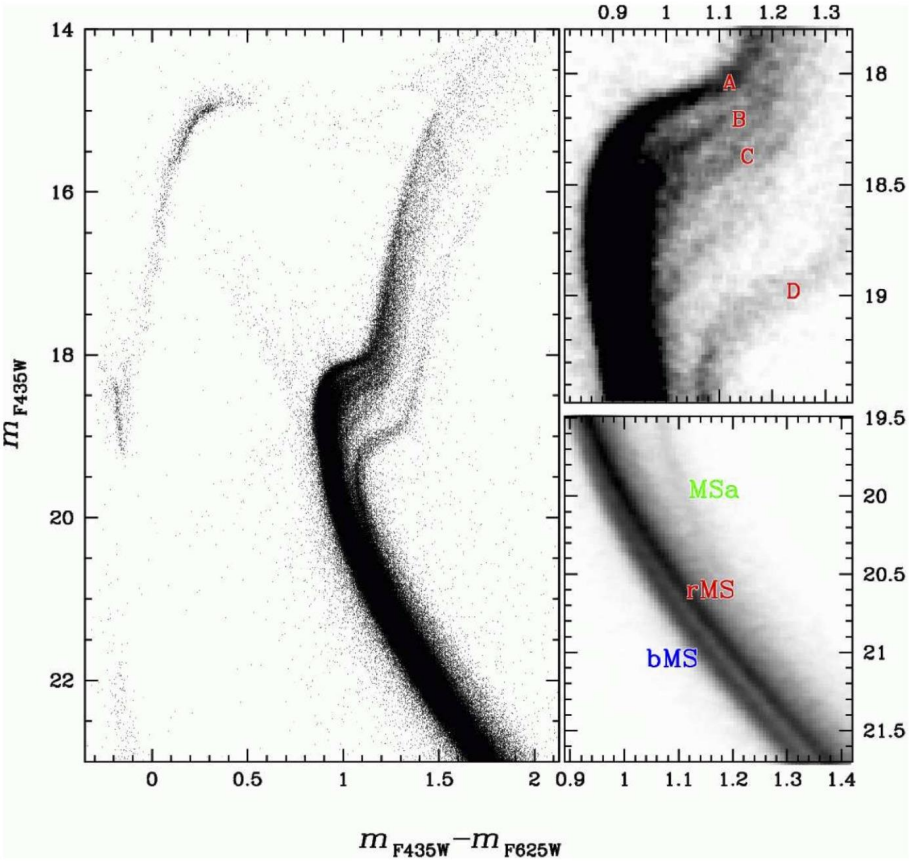

Figure 13: A color-magnitude diagram of the Omega Cen globular cluster reveals the presence of stellar populations with significantly different ages (as indicated by the tracks labeled A, B, C, D in the plot in the upper right). These data are from Bellini et al. (2010).

4.2.1.1 The Big Data Challenge: Resolved Stellar Populations

A number of steps are required for the study of resolved stellar populations, as outlined in Figure 15. One of the common, and most time-consuming, steps (#3 in Figure 15) is the determination of completeness corrections. Artificially adding stars to the image with a wide range of magnitudes and colors, and determining which can be recovered using the star finding algorithms, allow us to estimate fraction of stars that are recovered in the photometric measurements as a function of stellar type and brightness. As this can often be one of the more computationally intensive steps in studying stellar populations, we will use it to demonstrate the computing requirements for a state-of-the-art project; the 828-orbit *Hubble* multi-cycle treasury program known as the Panchromatic Hubble Andromeda Treasury (PHAT) project (Dalcanton et al. 2012; Williams et al. 2014).

The observations consisted of six *HST* observations in each of 411 fields. The PHAT team had used a 24 CPU core cluster for previous resolved stellar population surveys. However, to obtain the necessary photometric measurements and artificial star tests for the PHAT survey, each field required about 5,000 CPU hours to perform. Multiplying all this out we calculate that the total amount of time required using just the 24 CPU core system would have been 5,000 hours x 411 fields / 24 CPUs / 24 hour = 3,570 days ≈ 10 years! Instead, the PHAT team acquired the necessary computing resources for their survey by utilizing cloud computing. With 1,000 cores, their analysis required about 3 months. Looking forward, *WFIRST* will survey in just 2 pointings the same area of sky the entire PHAT survey covered with 411 pointings of *HST*. ***Hence, for the future, a 10K CPU core cluster will likely be needed. Such a system could analyze the PHAT data in about one week***.

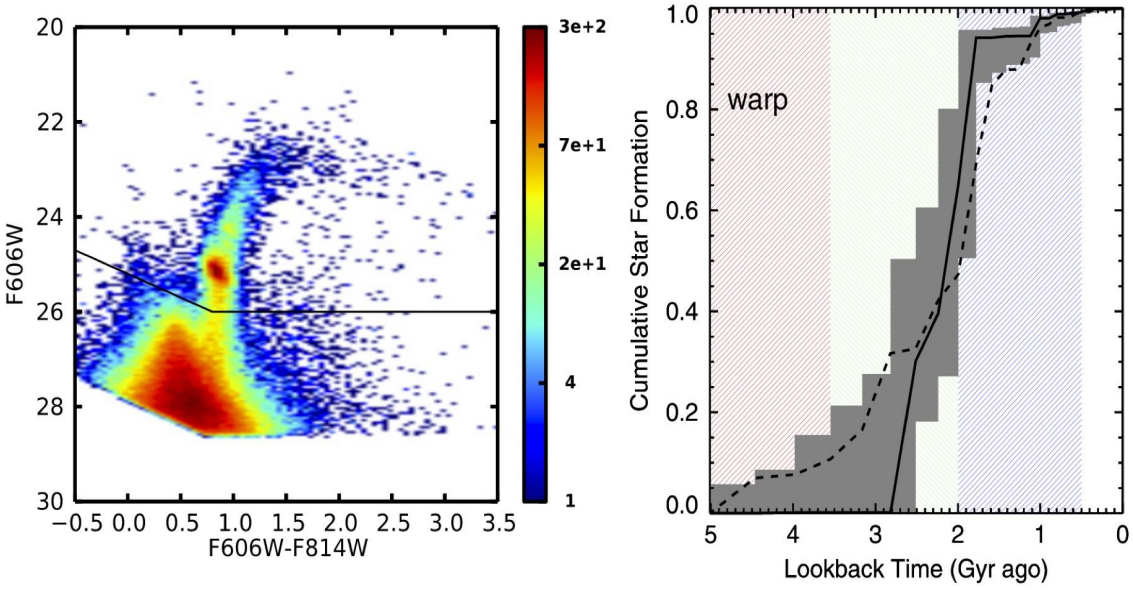

Figure 14: The color-magnitude diagram (left) and the corresponding star formation history (right) in the nearby galaxy, M31 (see Williams et al. 2015).

Storage and bandwidth requirements are generally less demanding for these types of projects. The ~2,500 Hubble images that comprise the PHAT program (~300 MB each) require about 1 TB storage. A network bandwidth of ~1 Gbps would allow one to transfer the images to the work site for the completeness testing in about 2 hours. Some of the other quick-look analysis/visualization tools may also require adequate bandwidth to be effective, but the more intensive steps could be reserved for data-host-side computing instead of being served remotely. While a 1 Gbps external network bandwidth is acceptable for many of the stellar population programs of today, ***a 10 Gbps external network bandwidth will be essential in the WFIRST era, when a dozen PHAT-sized surveys can be completed every day!***

There now exists within the astronomical community a large and continuously growing archive of imaging of resolved stellar populations from systems within the local Universe.

This will not only continue, but also accelerate with the upcoming *JWST* and *WFIRST* missions and the PanSTARRS and LSST surveys. The vast majority of existing observations were obtained in fairly small programs, which targeted one to a few galaxies or Galactic clusters, with the consequence that the published analyses (for example, star formation histories) cannot easily be compared. ***This precludes holistic studies of the star formation history of large numbers of galaxies in the nearby universe.*** Even with the current size of the resolved stellar population archive from *HST* alone, performing these studies requires a major investment in time and resources, in both computing and personnel. Initial attempts to do this have been very fruitful (e.g., ANGST and ANGRRR; Dalcanton et al. 2009) but are already quite out of date (ANGRRR and ANGST only considered *HST* photometry obtained before Cycle 17). The addition of *JWST* and *WFIRST* data will quickly make this an intractable problem for all but very large, very well-funded teams. ***The computer resources described here would make similar studies feasible for the community at large, which will undoubtedly greatly enhance the science that can be done with these datasets.***

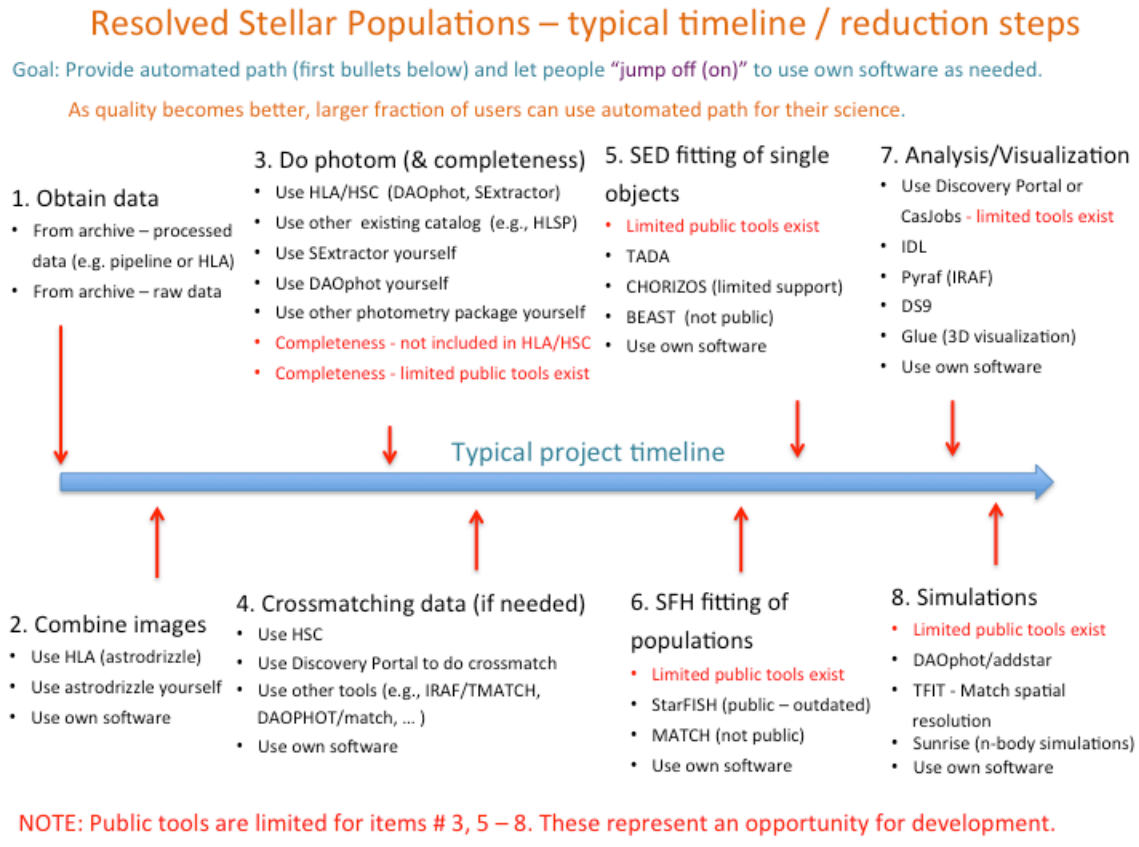

Figure 15: A workflow for a stellar population analysis. While automated steps using the *Hubble* Legacy Archive (HLA) and the *Hubble* Source Catalog (HSC) exist for the first four steps, but no public tools are available for the last four. Other commonly used tools are also listed.

Figure 15 can also be used to demonstrate a key opportunity that exists in the study of resolved stellar populations; namely the need for publicly available tools for several of the steps. While several groups worldwide have developed the necessary tools, this leaves a multitude of other astronomers that might want to do similar work on their own. Hence, the ***development of publicly available software*** for certain steps of the process that do not currently exist (e.g., Step 3 - completeness tests, Step 3 – Automated PSF-fitting photometry, Step 5 - SED fitting, Step 6 - SFH fitting, Step 7 - Analysis and visualization, Step 8 - Simulations), ***would be welcomed by the community and should be a strategic initiative for STScI***.

This approach – the development of an automated path at a basic level through the full timeline – is part of the long-term plan for the *Hubble* Legacy Archive (HLA) and the associated *Hubble* Source Catalog (HSC), as outlined in Figure 15. Further integration of this facility with other databases, both present (e.g., SDSS, *Spitzer*, *Chandra*, PanSTARRS) and future (*JWST*, LSST, *WFIRST*) would leverage both the software and hardware investments that are currently going into these projects.

4.3 Mapping the Cosmos in 3D

Matter is not distributed uniformly in the Universe. Galaxies are often found in groups, clusters and superclusters. Clusters of galaxies form from the collapse of baryons and dark matter halos over regions that span several million parsecs[6]. These structures are pulled together by gravity and lie at the intersections of filaments of dark matter that trace primordial density fluctuations in the Universe. Figure 16 shows a spectacular map of this "cosmic web" of structure as revealed from the SDSS galaxy redshift survey data. A key element in understanding the cosmic history of structure formation is the distances to the structures we see – be they galaxies, galaxy clusters, or superclusters. In a cosmological context, the distance not only gives us a three dimensional view of how matter is distributed but also provides the cosmic timeframe in which the structures exist.

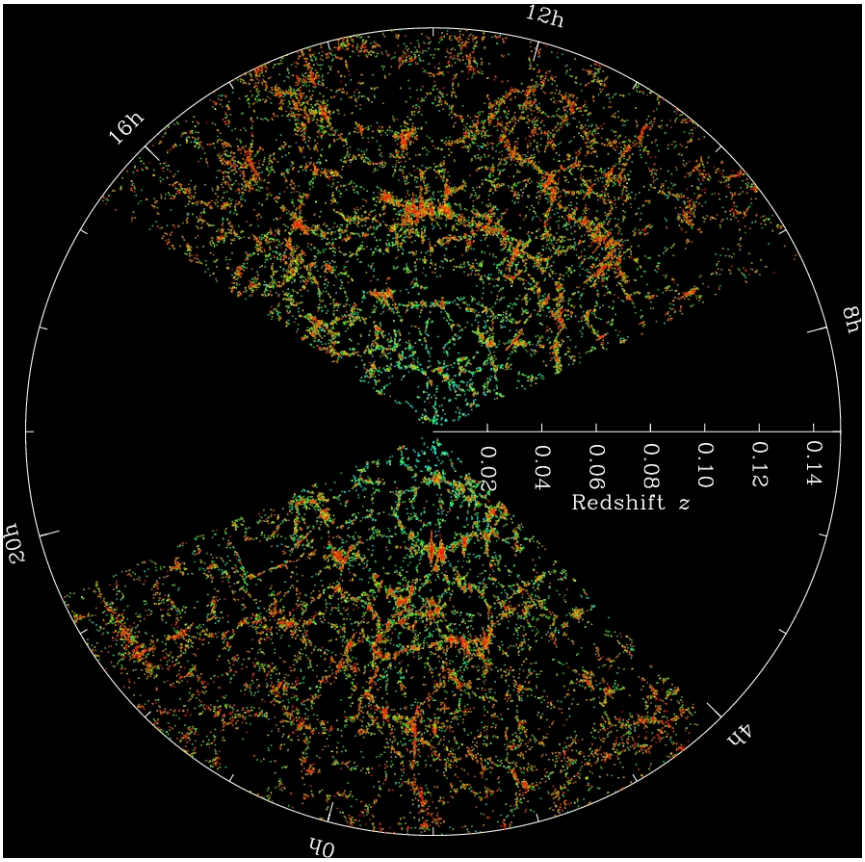

Figure 16: A quasi-3D projection of the distribution of galaxies in the Sloan Digital Sky Survey (SDSS). Redshifts are used as the primary galaxy distance indicator in this map. The map shows a region of space that is ~620 Megaparsecs (Mpc) in radius.

[6] One parsec equals 3.26 light years.

Multi-wavelength observations of high-redshift clusters (*e.g.,* Rosati, Borgani & Norman 2002, Williamson et al. 2011, Marriage et al. 2011) are particularly important for studies of evolution as they provide important measures of the dynamical evolution of these structures, on the epoch of formation of the most massive galaxies (which preferentially reside in massive clusters), and also provide constraints on the cosmological parameters.

The redshift of a galaxy indicates how fast the galaxy is receding from the observer due to the expansion of the universe. For galaxies beyond our Local Supercluster (*i.e.,* beyond a distance of about 20 Mpc or so), the redshift is also a very accurate indicator of how far away a galaxy is. Spectrographic redshifts are the most accurate but often there are either too many galaxies in a survey to obtain spectra for all, or in the case of more distant structures the galaxies are too faint to have their redshifts measured with a spectrograph, even with the largest existing telescopes. Fortunately, there is a technique that can provide a reasonable estimate of an object's redshift without relying on spectrographic data and only requires broadband photometry spread over 5 or more passbands. Redshifts obtained in this way are called photometric redshifts and can be generated for very large samples of objects, including objects that would otherwise be too faint to observe spectroscopically. As such, photometric redshift estimation is a fundamental tool to map the 3D distribution of galaxies in large area sky surveys. ***Over the next two decades, space-based survey telescopes like Euclid, WFIRST and wide-field ground based telescopes (e.g., PanSTARRS, VISTA, LSST) will open new opportunities to map cosmic structure on large scales***. The data collected by these surveys, in conjunction with databases provided by *HST*, *Spitzer*, *GALEX*, *JWST*, and other facilities, will allow us to derive accurate spectral energy distributions (SEDs) of galaxies over an unprecedented range of wavelengths.

4.3.1.1 The Big Data Challenge: On-demand photometric redshift mapping

The amount of data collected in such databases requires a new approach on how to process and distribute photometric redshift information to the astronomical community. As the sky becomes more thoroughly covered with multi-wavelength data, tools that enable mapping the 3D structure of the cosmos by estimating photometric redshifts of galaxies will be in high demand. Such tools could, in principle, be used to map structure in a narrow beam and, with proper computational scaling, can be extended to map the full area covered by each of the surveys used. Even for relatively small fields (*e.g.,* a single exposure from *JWST* spans just a few square arcminutes) this may require the simultaneous analysis of tens of thousands of objects, resulting in significant CPU time with current computers.

The procedure to derive photometric redshifts of galaxies in a survey involves a series of steps. The basic inputs are the celestial coordinates and field radius of a given patch of sky. The user would also want the ability to select the catalogs and observing bands to be included in the measurement of the photometric redshift and the algorithms to be used

(where different options are available). We describe the main steps in the photometric redshift estimation process below.

4.3.1.2 Galaxy Identification

The first step is to identify, in each passband, the source that corresponds to each individual galaxy. This is essential to minimizing the errors in the SED fitting and photometric redshift estimation. This is also an extremely difficult task. Machine learning techniques could be developed to be able to deal with complex objects in order to disentangle the correct object at different wavelengths. In Figure 17 we show an example from the COSMOS survey, in which the object of interest (which is also a radio source) is undetected at wavelengths blueward of the optical z band (~850 nm wavelength). In the IR, however, the target becomes significantly brighter than the other nearby objects. The photometric redshift estimate for this galaxy using *SExtractor* processing performed without careful attention to the identification of the correct object in each band leads to a significant underestimate of the redshift (z = 1.2 versus the accurate value of z = 1.9) and a larger uncertainty. Steps to address this challenging problem include: (1) measuring the flux in all bands within the area covered by the object in the band where it has the largest apparent size (this is feasible only if all observations have roughly similar angular resolution, *e.g. HST* WFC3-UVIS and WFC3-IR, *HST* and *WFIRST*); (2) measure each component detected in the highest resolution image, and decide, *a posteriori,* which components should be merged to derive the SED of the target (this could be useful if the goal is to select Lyman break galaxies); or (3) adopt some optimal mix of these two approaches, if the analysis warrants it.

4.3.1.3 SED-Fitting and Photometric Redshift Estimation

For each object, a best-fit to the SED is derived from a linear combination of a suite of galaxy and AGN spectral templates, or modeled via machine learning techniques. Ideally, the outputs are a probability associated with each model, a redshift probability distribution, and best fit solution with statistical errors. There are a number of software packages already available that do precisely this. A few examples are Hyper-z (Bolzonella et al. 2000), BPZ (Benitez et al. 2000, Coe et al. 2006) and EAZY (Brammer et al. 2008). These methods are all based on template fitting. Dust emission and dust reddening must, in general, be included in the fitting process. Recently, Beck et al. (2016) have developed a machine learning technique to empirically measure SEDs and photometric redshifts in the SDSS. In this method, selection of a reliable reference training set is required.

4.3.1.4 Detecting Large-Scale Structures

Once the photometric redshifts are obtained, the last step is to statistically search for large-scale structures, *e.g.,* overdensities of galaxies in a quasi-three-dimensional volume. We use the prefix *quasi* here because, even in surveys with many photometric bands, the errors in an individual photometric redshift can be comparable to the size scale of some of the

structures one might want to detect. Hence, for robust large-scale structure (LSS) detection, one will need hundreds, if not thousands, of galaxies to sample the structures that span the image or suite of images being analyzed.

Detection Algorithms and Output Products: Several LSS detection algorithms have existed for many years. For example, the "friends-of friends" method is based on finding galaxies pairs that are closer to another pair than a given redshift cut-off separation (Huchra & Geller 1982; Botzler et al. 2004). The COSMOS survey used a method based on 3D adaptive smoothing to highlight overdensities in a low S/N regime. Other methods use wavelet algorithms. The PPM (Poisson Probability Method, Castignani et al 2014a,b) was recently derived to find clusters of galaxies using photometric redshifts in pencil beam regions. All of these methods require significant CPU time and must be optimized possibly using parallel programming and GPUs.

One output product would be a 3D map of the objects in the region of the sky initially chosen by the user. Galaxy identification and SED/photometric redshift fitting may be performed each time, or a database with precompiled results could be created by the STScI and periodically updated with the best photometric redshift templates or training sets as new or better data become available. The downside of creating a static database is that the user may want to include or exclude specific instruments or passbands, depending on the goals of the research or the specific redshift range being studied. For example, some low-S/N medium and narrow band photometry might be useful if the user is mapping the distribution of AGNs but for those searching for distant galaxy clusters, the inclusion of those data may reduce the accuracy of the photometric redshifts for the cluster galaxy members.

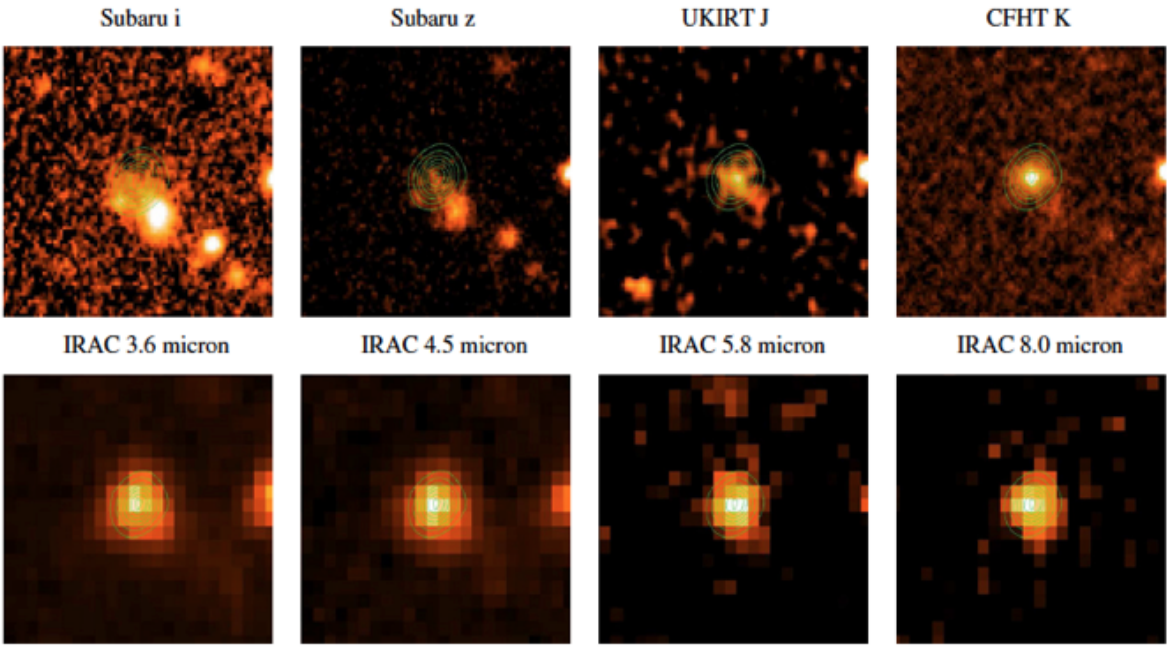

Figure 17: A multi-wavelength view of a distant galaxy (from Baldi et al. 2013). Superposed on the optical and infrared images are the contours (green lines) of the detected source in radio wavelengths. Standard object detection procedures often fail to identify the correct optical or IR counterpart to the radio source. In this particular case, the galaxy associated with the radio source is only detected at wavelengths longer that ~850 nm. Inclusion of incorrect source data leads to a significant error in the photometric redshift estimate.

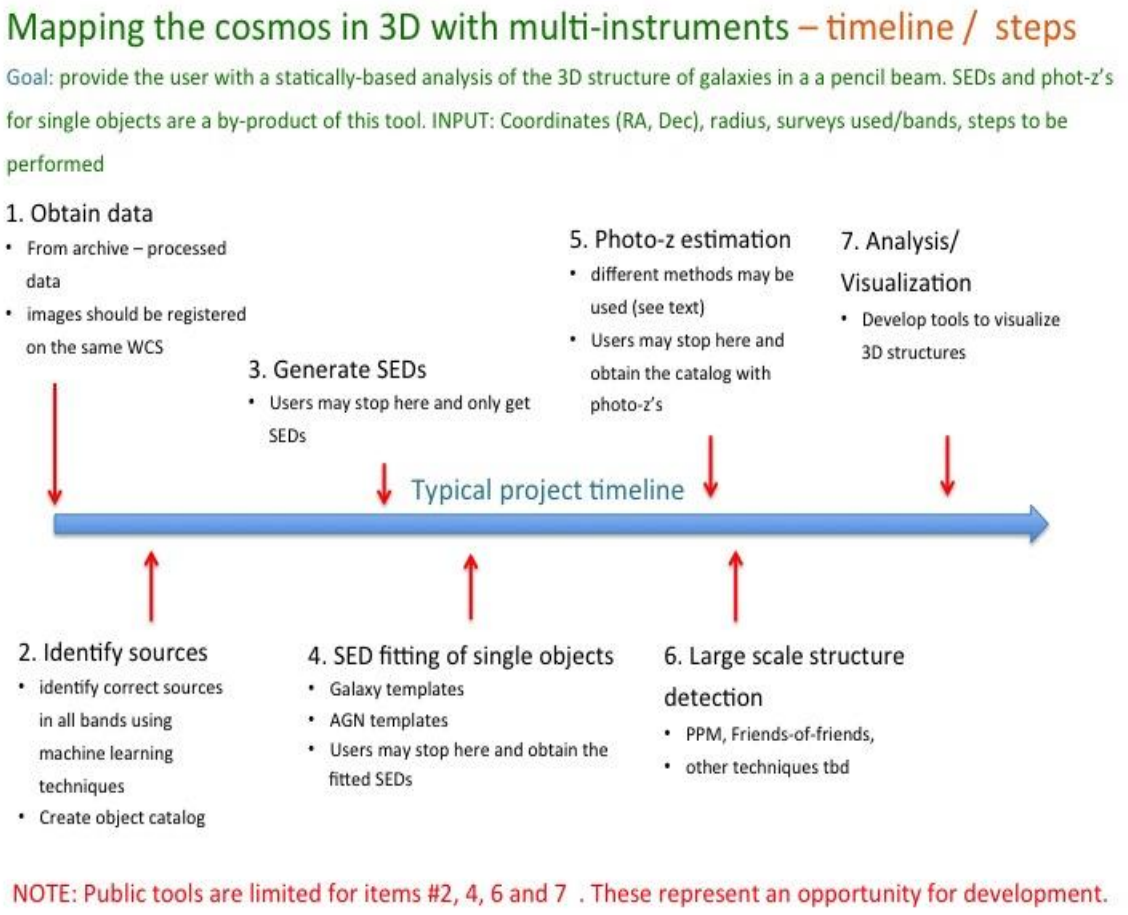

Figure 18: Workflow for the measurement of photometric redshifts and the subsequent detection of large-scale structure in a given region of the sky.

Computing and Storage Requirements: In order to perform all of the tasks outlined required for LSS detection (see Figure 18) from galaxy identification in each passband to the detection of the 3-D structures, ***it is highly desirable to have as much of the data stored at the same location as the computers used for the analysis***. This argues in favor of server-side deployment of the photometric redshift and structure finding algorithms as one will want to perform such analyses on a union of many survey datasets and, hence, processing a large number of images from each survey may well be required. For example, we estimate that about 160 TB of storage will be needed for the 2500 deg^2 *WFIRST* high-latitude survey images (4 filters), assuming that the drizzled pixel scale if the final image products is 0.06"/pixel. To cover the same area of the sky with LSST data in six bands, one would need an additional ~24TB of disk space for the reduced images. The galaxy identification step, which produces the main catalog, can be done once, or periodically using newer/deeper high-level science products from the surveys.

The primary computing constraints will come from the need to run object detection on large numbers of images and then cross-match potentially very large catalogs across a range of wavelengths. Both existing and planned wide-field sky surveys will contain between 10^7 to several 10^9 galaxies. Efficient methods, built for SDSS, can be used to perform such catalog cross-matching in a matter of minutes on optimized systems with a few hundred CPU cores. For smaller regions of the sky (as defined by the user through a web page), a pre-prepared source catalog may be the most efficient option. For mapping projects that might only be studying regions of ~10 Mpc in scale, photometric redshifts and LSS detection may only need to process a sample of about 100,000 objects. This can be done with optimized codes running on a high-performance (multi-core) desktop in a relatively short time (~1h or less).

4.4 Black Hole and Host Galaxy Co-Evolution

Most galaxies appear to host a large central supermassive black hole at their core, with these black holes typically being a few million to a few billion times more massive than our sun. There is strong evidence of coupling between the properties of these black holes and their host galaxies, as indicated for example by observed correlations between the central black hole mass and the mass of the central bulge of galaxies (Ferrarese & Merritt 2000 and Gebhart et al. 2000; Gultekin et al. 2009; DeGraf et al. 2015). However the physical processes that determine this interplay between black hole and galaxy properties remain largely unresolved.

One way to probe these relationships in more detail is to study galaxies that contain an "active nucleus". In these sources, the central supermassive black holes are thought to be surrounded by disks of accreting gas (Rees 1984; Ho 2008), which heat up as a result of the intense pressure and produce emission that peaks at extreme ultraviolet and X-ray energies. If the accretion rates are sufficiently high, the resulting energetics can also drive powerful outflows or winds from the central regions, which can in turn impact the rest of the galaxy in "feedback" effects (Springel et al. 2005). In the most extreme cases, the outflows can be powerful enough to clear out all the gas from the central regions, and which may significantly disrupt star formation in the central portions of the galaxy (di Matteo et al 2005, Hopkins et al. 2009).

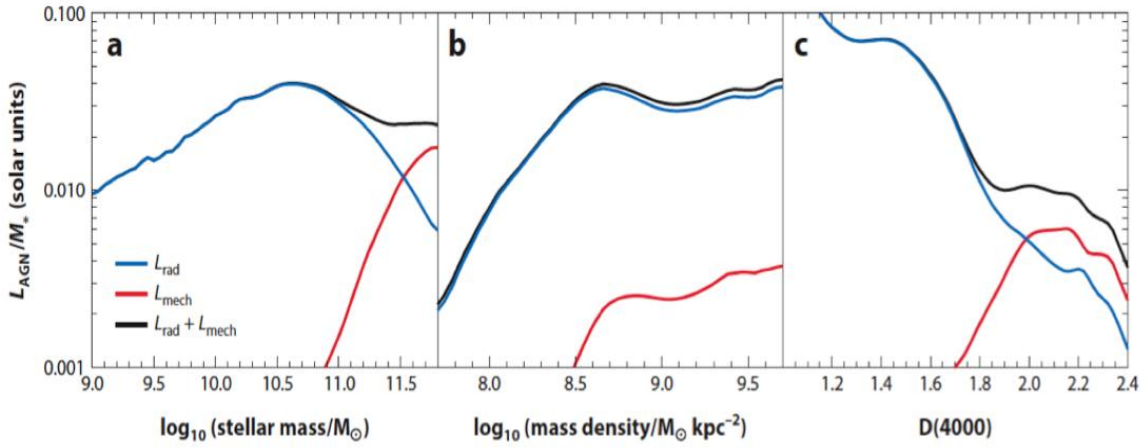

Figure 19: Results from the SDSS (Heckman & Best 2014) showing how the dominant energy output from the central black hole changes from radiative mode at the low mass end to kinetic output (*i.e.,* jets) at the high-mass end. The overall dominance of the luminosity of the active galactic nucleus relative to that of the host galaxy is highest in galaxies with younger ages (i.e., systems with D4000 < 1.5 in the rightmost plot; D4000 is an age-sensitive feature in the spectra of galaxies).

Results from the Sloan Digital Sky Survey (SDSS) by Heckman & Best (2014) indicate that, at least for galaxies in the nearby universe, the energy output mode of active galactic nuclei appears to be related to the mass of the host galaxy. For example, in low-mass galaxies the

energy output seems to be primarily "radiative", consisting of ultraviolet and X-ray emission. On the other hand, galaxies with higher mass appear capable of ejecting powerful outflows or "jets" of material, in which case the energy output is primarily in the form of kinetic energy (Figure 19). In addition, "feedback" effects, where the outflows directly impact the surrounding galaxy, are directly observed in a number of these sources, but it remains uncertain how significant these effects are in regulating the overall star formation rate and subsequent growth of these galaxies.

A primary limitation to our quest to understand the black hole – galaxy connection has been the relatively small sample sizes of active galactic nuclei that can be studied in sufficient detail, especially considering the need to obtain high resolution imaging on their host galaxies, which currently is obtainable only using space-based telescopes such as the *Hubble Space Telescope*. As a result, active galaxy samples with well-resolved morphologies have generally been limited to less than a few thousand objects in total, across the full range of luminosity, host galaxy properties, and accessible redshifts. Moreover, the SDSS studies discussed above, while covering a larger area, are typically limited to much shallower depths and are therefore relatively complete only in the local universe. As a result, inferences drawn about the importance of feedback processes from black holes on their host galaxies tend to be limited to the specific details of the samples being considered, which are often too small to allow sufficient sub-division across parameters of interest (*e.g.*, luminosity, mass, galaxy morphological properties, redshift, or galaxy environment), and are not generalizable to the full population.

The new generation of deep all-sky surveys coming on-line in the near future (*e.g.,* PanSTARRS and LSST from the ground, *WFIRST* from space yielding *HST*-class morphologies, and all-sky multi-wavelength surveys from facilities such as *eRosita* at X-ray wavelengths and the Square Kilometer Array [SKA] at radio wavelengths), ***can be expected to increase the number of known active galactic nuclei by several orders of magnitude***. For example, the *WFIRST* High-Latitude Survey (HLS) would cover ~2500 deg^2 and yield *HST*-quality morphological information on up to ~400 million galaxies, likely containing up to several million active galactic nuclei. Moreover, LSST is expected to deliver ground-based imaging of up to ~10 billion galaxies, with likely up to several hundred million active galactic nuclei detected by means of time-variability.

These datasets will for the first time enable black hole / host galaxy properties to be studied for the full population of galaxies across much of cosmic time, thereby finally removing small sample size as one of the primary limitations to progress. However, a key challenge in making effective use of these datasets is development of sufficiently capable tools to extract the relevant information from the rich datasets that will be available. As an example, *HST*/*WFIRST* can provide many pixels across a typical galaxy (*e.g.*, a galaxy ~1" in diameter is sampled by ~100 resolution elements in a single band) and reveal complex morphological information not adequately captured by simple parametric models, as illustrated in Figure 20.

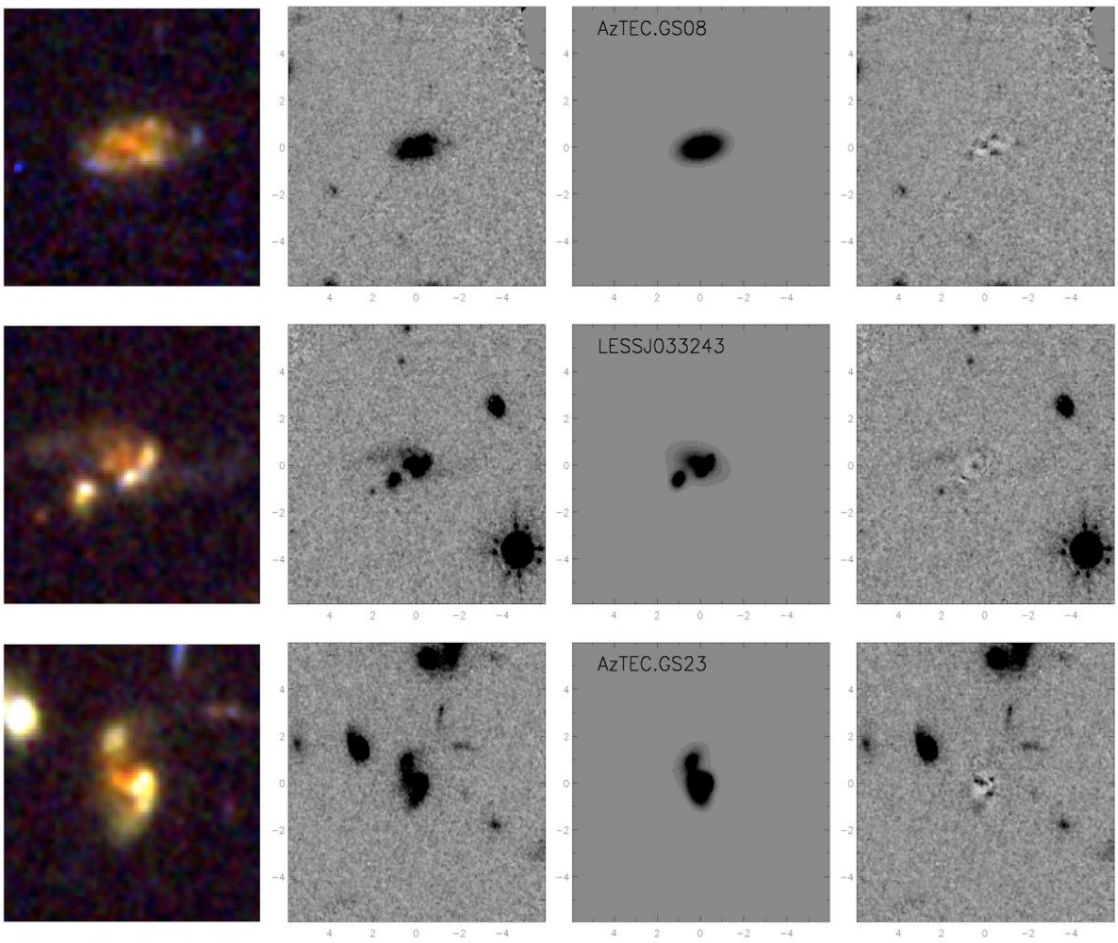

Figure 20: Multi-band *HST* imaging of galaxies from CANDELS (Grogin et al. 2011, Koekemoer et al. 2011) for a sample selected from Targett et al. (2013) that cover a representative range of complex morphologies. Note especially that the simple parametric fits (third column of images) do not adequately capture the full morphological information, as evidenced by the complex central structures remaining in the residual images (rightmost column).

Specifically, the following workflow summarizes the capabilities and tools that would be needed in order to accomplish the goals of extracting the relevant information from these new samples:

1. ***Development of morphological descriptors*** that are able to capture the full multi-band information present in at least ~100 resolution elements per galaxy that will be expected from datasets obtained with *WFIRST*, ***extending beyond the current morphological classification paradigms*** that are really only applicable to galaxies with a few resolution elements at most (since most galaxies beyond $z > 0.5$ are barely resolved in ground-based data).

2. ***Efficient ways to cross correlate the resulting detailed multi-wavelength morphological information*** (potentially up to ~100 quantities per galaxy, for up to ~10^{10} galaxies), in a way that is computationally feasible.
3. ***Effective way to visualize the resulting N-dimensional parameter space*** (with N being large, potentially up to ~100, populated by up to ~10^{10} galaxies).
4. Comparison with results from simulations, in a similar N-dimensional space, to enable full exploration of the various scenarios for how AGN and their hosts affect each other's evolution.

4.4.1.1 The Big Data Challenge: Galaxy – Black Hole Co-evolution

The primary driver for requirements in terms of disk space is the need to store all the image data that is necessary for carrying out the morphological analysis. For the *WFIRST* HLS imaging data, we expect to start from the combined full-depth mosaics, assuming a pixel scale of 0.06"/pixel for final combined images (*i.e.*, half the *WFIRST* detector pixel size, taking advantage of sub-pixel sampling from appropriately dithered input exposures). Given the total area of ~2,500 deg^2 for this survey, this corresponds to ~40 TB, at 4 bytes per pixel, for each type of image product (with the 3 likely types of products needed being science, weight and exposure time images), thus 120 TB per filter, or 480 TB required for the full set of 4 filters for this survey. Similarly, the full-depth images from the LSST main survey, consisting of 18,000 deg^2 covered with 0.2" pixels in 6 filters, would correspond to 24 TB for each of the 3 types of image products, thus 432 TB for all the image products from all 6 filters that are used this survey. The total disk storage requirement is therefore just over 1 PB for the full-depth image products and ancillary files from the *WFIRST* HLS survey and the LSST main survey.

The requirements on processing speed are derived from current experience in running morphological analysis software in large *HST* projects such as CANDELS, COSMOS and others. For the *WFIRST* HLS data, where the pixel scales are directly comparable to those on *HST*, we can use our existing experience of ~1 minute required, per galaxy, for morphological analysis software running on a single current state-of-the art CPU. Since each galaxy is run independently, this means that this process can be fully parallelized, thereby enabling us to obtain morphological measurements on the full set of ~4x10^8 galaxies in the HLS survey over a period of ~4 months using a cluster of ~2000 CPU cores (with the time required being directly proportional to the inverse of the number of CPU cores). Since the LSST pixel scale is coarser, at 0.2"/pixels (thus 11 times larger in area than the 0.06" pixels in *WFIRST* HLS co-added images), the time per galaxy in LSST data would be reduced correspondingly; running the same software on the ~10^{10} galaxies in LSST could therefore be achieved in ~10 months using the same ~2000 CPU core system (with the required time again scaling inversely with the number of CPU cores).

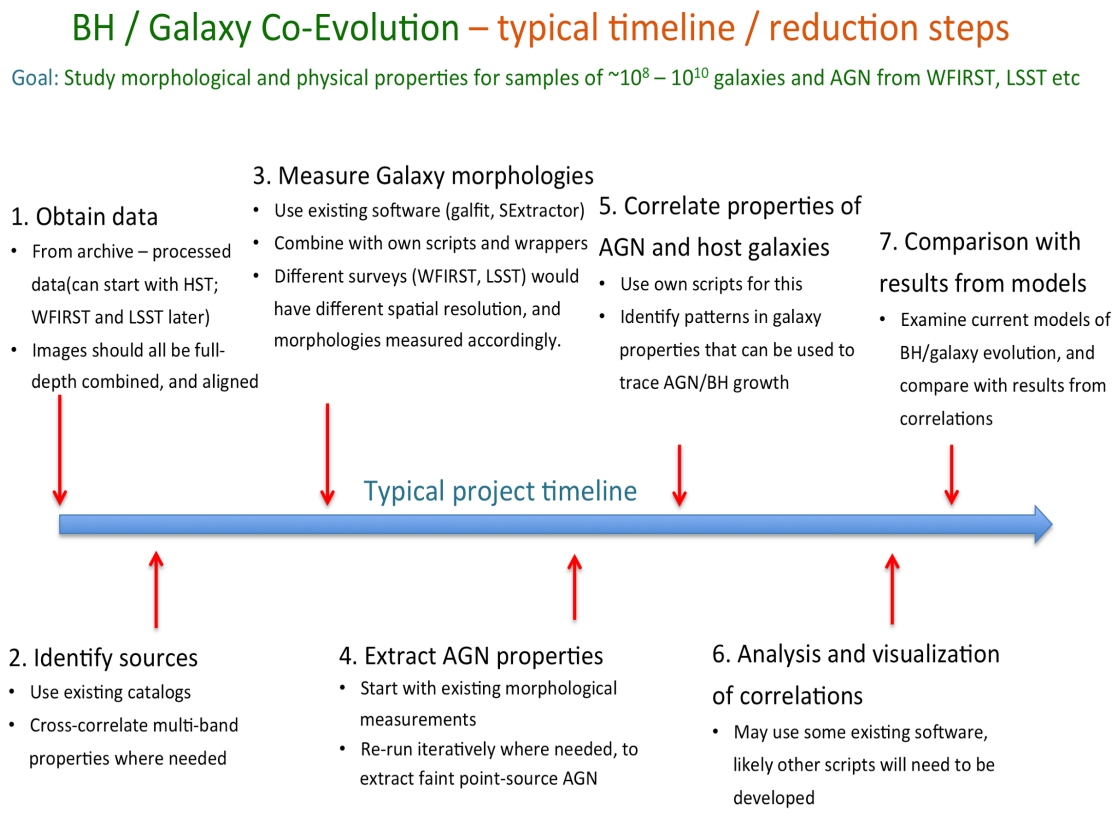

Figure 21: Workflow for the measurement of galaxy and black hole properties from imaging data obtained initially from *HST* and, eventually, from *JWST*, LSST, *WFIRST* and other surveys.

A successful approach for breaking down the above workflow into a 5-year timeframe (see Figure 21) would be to focus during the initial years on existing data from *HST* and PanSTARRS as a starting point for setting up a database of integrated photometric information on the full set of galaxies that are available, which could be cross-correlated with existing all-sky surveys at other wavelengths. In parallel, development of tools to quantify more fully the rich morphological information present in *HST* images could be started, which should also allow measurement of structural properties across multiple bands. Once data from LSST and *WFIRST* become available, along with large-area surveys from SKA and *eRosita*, these tools could then be extended to accommodate the subsequent increase in active galaxy sample size, along with morphological information on their host galaxies. For the first time ever, we would then have the ability to study the properties of the full population of galaxies and their associated central black holes, and can finally hope to resolve the nature of the physical mechanisms that drive their co-evolution.

4.5 Science with Time Domain Surveys

4.5.1 Constraining the SN Ia Progenitors with High-Cadence *TESS* Light Curves

The measurement of the expansion of the universe requires access to precise distance estimates to galaxies across a vast range of cosmic time. Type Ia supernovae (SNe) have been found to be one of the most precise cosmological distance indicators. Type Ia SNe are believed to occur in binary star systems in which at least one of the stars is a white dwarf. And yet, although this type of SNe has been essential in helping us make profound discoveries (*e.g.,* the accelerating universe), the progenitors of Type Ia SNe remain a mystery, limiting our physical understanding of Type Ia SNe. We still do not know if Type Ia supernovae explosions come from single-degenerate[7] binary stars (*e.g.,* only one of the stellar companions is white dwarf star) or binaries composed of two white dwarfs (*a.k.a.* a double-degenerate binary). Observations of a supernova immediately following the explosion provide unique information on the distribution of ejected material and the progenitor system.

Recent theoretical models show that the initial behavior of the light curve (the brightness of the supernova over time) is quite different for the two possible origins of Type Ia SNe: the secondary star in a single-degenerate binary will cause bright shock-wave induced emission in the first hours or days after the explosion, while double-degenerate explosions are expected to brighten monotonically (Kasen et al. 2010, Piro et al. 2013, see left panel of Figure 22). However, to observe a supernova immediately after detonation requires continuous monitoring of many galaxies. Previously, only a handful of very nearby SNe Ia have been observed during the first few days. The SDSS-II survey contains one of the largest samples of early SN Ia light curves (Hayden et al. 2010a), but the survey had an average cadence ***exceeding four days***. No clear signature of companion interaction was found in the SDSS-II SN Survey (Hayden et al. 2010b), ruling out red giants or larger companions. Similar null results were found in analyses of 87 SNe from the Supernova Legacy Survey (Bianco et al. 2011), and 61 SNe from the Lick Observatory Supernova Search (Ganeshalingam et al. 2011).

With the *Kepler* space telescope, we are now able to obtain early light curves of supernovae in unprecedented detail owing to its observation cadence of 30 minutes, well within the required timescale to differentiate between a single-degenerate or double-degenerate progenitor. The three SN Ia found in a survey of 500 galaxies with *Kepler* between 2010 and 2012 provide the best rise-time information ever obtained for thermonuclear SNe (Olling et al. 2015). See the righthand panel of Figure 22 for an example. These SN Ia light curves are all well fit by a single power law, but they found no signatures of the supernova ejecta

[7] A binary star with at least one white dwarf is referred to as a "degenerate" system because, unlike main sequence stars that balance gravity's inward pull by thermonuclear fusion in their cores, white dwarfs are supported by electron degeneracy pressure – a quantum mechanical effect.

interacting with nearby companions. With the extended operations of the degraded *Kepler* mission, called *K2*, the number of SN Ia with excellent space-based light curves will increase to about a dozen over the next couple of years.

The *Transiting Exoplanet Survey Satellite* (*TESS*), a successor to the *Kepler* mission, will transform this set of high-cadence SN Ia into a statistically significant sample. Even though the aperture of *TESS* is 10 times smaller than that of *Kepler*, the rate of observed SN Ia will be larger since the full-frame images (each FFI covers ~2300 deg^2) will be regularly downloaded, rather than just downloading postage stamps centered on pre-selected targets as is done with *Kepler*. This is a real paradigm shift, since the *TESS* data stream will be more like ground-based transient surveys where it is necessary to identify a relatively small number of variables and transients within a large dataset.

In further contrast to the *Kepler* mission, most of the *TESS* fields will be observed for a short period of time (28 days). Since SNe have time scales of several months, only part of their light curves will be covered with the *TESS* observations and that is not enough to determine the supernova type solely from the light-curve data[8]. To optimize the SNe discovery potential of *TESS*, one would require a *TESS* operations plan that enables timely discovery of the SNe in its FFI data to allow for rapid spectroscopic follow-up for classification from the ground. This can be achieved by either analyzing the TESS data within a few days of the downlink to the ground-station, or by initiating a ground-based imaging campaign that simultaneously covers each of the 2,300 deg^2 *TESS* fields. If either of these options is realized, the *TESS* datasets could be a treasure trove for the study of supernovae.

4.5.1.1 The Big Data Challenge: Finding Supernovae in TESS Full-frame Images

TESS has four detectors, each 4K x 4K in size spanning a field-of-view of 576 deg^2. The full-frame images of all detectors are read out every 30 minutes, leading to a modest data rate of 6.5 GB/day. Over the full 2-year lifetime of the mission, the total raw data volume is thus on the order of 6.5 TB, well within current capabilities. The bigger challenge, however, is reducing this dataset, especially if it is done in near real-time to discover the SNe. For comparison, the *Kepler* telescope only downlinked on the order of 20,000 cutouts around pre-defined objects. On timescales of minutes to hours, the photometric precision of this data are a few parts in 10^6. On longer timescales, this precision is considerably reduced due to many systematic effects introduced by the instrument. A specialized analysis is required to account for all of these effects (see Olling et al. 2015). In particular, a given epoch cannot be reduced independently of the other epochs, but the temporal evolution of instrument artifacts needs to be fit for all available epochs of a given cutout. The reduction of a typical *Kepler K2* cutout (22 x 22 pixels) takes on the order of 30 seconds.

[8] In other words, one could detect that a supernova went off but one could not determine whether it was a Type I or Type II. Without the supernova type information, the knowledge that can be gleaned from light curve data will be very limited.

We extrapolate the CPU requirements for *TESS* FFI processing based on *Kepler* data processing knowledge. The *TESS* Faint Catalog will contain about 470 million point sources and 125 million extended sources over the full sky. Therefore, a given 2300 deg^2 field will have on the order of 7 million extended sources. Since SNe can outshine their host galaxies, for the science goal of finding as many SNe as possible, all of these extended sources should be monitored, even if they don't pass the signal-to-noise threshold in a single FFI image. *TESS* downloads all new data every two weeks at its closest approach to Earth. This data needs to be reduced within a day from the time of the downlink in order to trigger ground-based observations like spectroscopy for newly discovered transients. Assuming a reduction time of 30 seconds for a single cutout similar to *Kepler*, about 2600 CPU days are needed to process the cutouts of all 7 million extended sources in the field. ***Therefore, thousands of CPUs are needed for a timely reduction of TESS FFIs to enable transient science***.

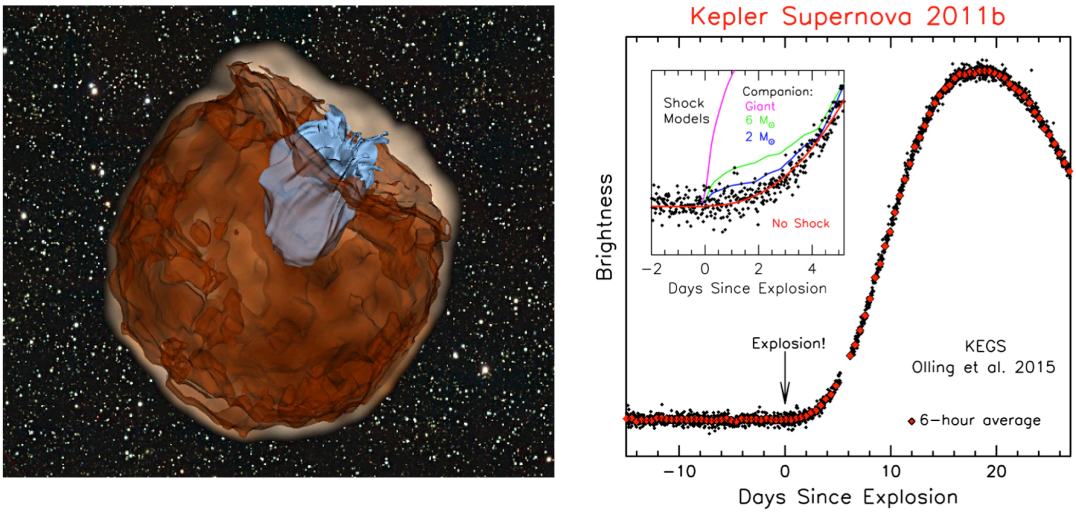

Figure 22: Left: A simulation of the expanding debris from a supernova explosion (shown in red) running over and shredding a nearby star (shown in blue). Image from Daniel Kasen (LBNL). Right: A light curve of a Type Ia supernova observed with the *Kepler* telescope. The filled black and red circles indicate the 30-minute and binned 12-hour data, respectively. Inset: Expected brightening if shock wave hits a companion star. No evidence of a companion was seen in this case.

We note that at present the *TESS* science operations plan may not be geared for rapid distribution of downlinked FFI data. This option should be explored, as the scientific payoff will be quite high.

4.5.2 The GALEX Inter-Visit Variability Catalog

Energetic phenomena in the universe often reveal themselves by sudden outbursts of energy in the ultraviolet (UV) wavelength regime. Study of such phenomena provides new insights into the physics of the cosmos. The *GALEX* mission conducted an all-sky survey in the UV, covering wavelengths from 134 – 283 nm in two broad passbands over its 10 year mission from 2003 to 2013. Each observation (what we refer to as a "visit") lasted anywhere from 60 seconds to 30 minutes. It is an ideal dataset to understand the energetic physics that influence star and galaxy evolution. To allow researchers to discover and characterize short-lived energetic events within the *GALEX* UV data, MAST recently released the gPhoton service. This service is composed of a 1.1-trillion-row, 130 TB database of nearly every *GALEX* photon event observed during the ten-year mission. A photon event includes information about the energy and arrival time of the photon on the *GALEX* detector array. The gPhoton service also includes open-source python software that can be used to create calibrated light curves and images at user-specified spatial and temporal scales. For the first time, MAST users can study *GALEX* UV variability at the inter-visit level (from seconds to minutes) without having to begin at the raw data stage.

With the gPhoton service users could construct a comprehensive catalog of variable sources, as revealed by significant changes in the brightness of astronomical objects during the course of a given *GALEX* observation (i.e., sources that change their luminosity over timescales of ~30 minutes). We refer to such sources as inter-visit *GALEX* variables. While some of these objects will also vary over longer timescales, a substantial fraction will be objects that vary largely, or exclusively, over these shorter time periods. These objects enable science across a diverse range of fields within stellar and galactic astrophysics, including stellar flares, white dwarf pulsators, and supernova shock breakouts.

4.5.2.1 Three Science Cases of Inter-Visit Variables

Low-mass stars (specifically, "M dwarf" stars that are less than half the mass of the Sun) are known to exhibit flaring events over a range of energies and occurrence rates. These flares are similar to those that occur on the Sun, and feature a sharp, sudden increase in flux, followed by a gradual return to pre-flare levels. They can range in duration from less than a minute to an hour or more, and can increase in flux by a few tens of percent to more than a factor of 100. Interest in M dwarfs as habitable planet host stars has steadily increased, but the question of how stellar flares impact the habitability of planets is also becoming increasingly relevant. Some work has been done studying the impact of the largest flare events these stars undergo, but a comparatively under-explored area of study examines the smaller flare events, specifically the flaring rate (how many flares per hour) and the energy distributions (counting how many flares on a given star increase its brightness by a given amount). These smaller, but potentially more frequent, flare events may have as significant an impact on habitability as the larger flares, but have been challenging to observe in large numbers given their very short durations (a few minutes) and unpredictable occurrence. The lower-limit of flare rates, even using *Kepler* data, occurs at flare energies (total energy

emitted by the flare itself) of less than 10^{29} ergs. Detecting and characterizing these smaller "micro" flares is optimized with gPhoton, since flare contrast is maximized in the UV, and *GALEX* has high time resolution and large spatial coverage across most of the sky (see Figure 23).

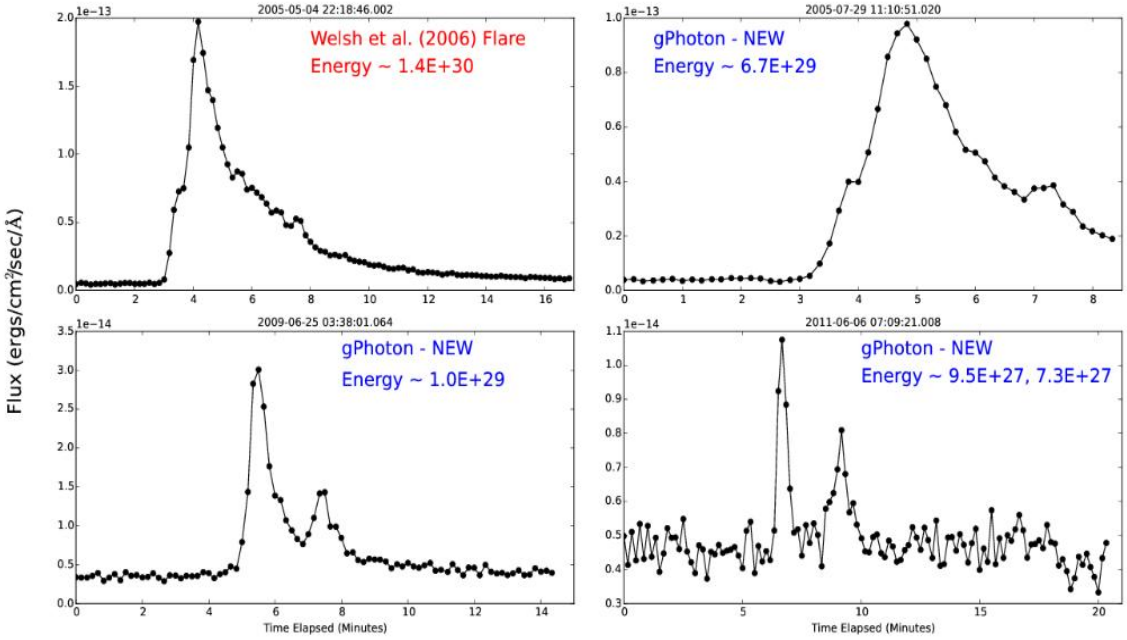

Figure 23: gPhoton flares in the NUV from CR Draconis, a well-known flaring M dwarf. The top-left flare was previously known, but in the full corpus of *GALEX* data there are at least 8 total flares, some with energies as low as 10^{27} ergs. The rates for these types of flares around stars other than the Sun are not well constrained and are precisely where the gPhoton data are optimized.

Another source of stellar inter-visit variability are white dwarfs (remnants of stars that have exhausted their fuel) that pulsate, changing their brightness over the course of ~1 to 30 minutes. These fluctuations in brightness are a few-percent if observed at optical wavelengths but can be up to ten times higher when observed in the UV. Because these changes occur over timescales of ~10 minutes, the gPhoton service is well suited to detect such pulsations within a typical *GALEX* observation. Critically, the ratio of the pulsation amplitudes between the UV and optical can be used to probe the interiors of these exotic stars and understand their compositions and the physics happening deep in their interiors. They can also be used to derive very precise masses and radii for the star, which can be compared with predictions made by stellar evolution models.

Much less common sources of inter-visit variability, but ones with great scientific potential, are supernovae shock breakouts. A shock breakout occurs when the shock wave, originating from the center of an exploding star, reaches the surface of the star's infalling atmosphere. This shock lasts only a few minutes in X-rays, but can last several hours in the UV. This brief phase of a supernova explosion is challenging to observe, since it happens a few days before the maximum light output in the optical is reached, but it is critically

important in understanding the origin of a supernova since the amplitude and duration of the shock breakout depends strongly on the size of the progenitor star. There is a lot of interest in determining how many supernovae are caused by single, large giant stars and how many are caused by a white dwarf accreting too much gas from a nearby stellar companion. Thus, observing and modeling shock breakouts is one way to determine the progenitors of different types of supernovae, and indeed, a handful of cases have been studied using *GALEX* and *Kepler*. The strength of the gPhoton service is that it converts any *GALEX* tile into a high definition light curve for any supernova breakout that happened to occur during the observation. Thus, the wide sky coverage of the *GALEX* survey enables a statistical study on supernova shock breakouts, despite their intrinsic rarity.

4.5.2.2 The Big Data Challenge: Creating the Inter-Visit Variability Catalog

Construction of the Inter-Visit Variability Catalog will follow in two stages (see Figure 24 for a representative work flow). The first is to analyze the light curves of previously identified sources from the mission's merged source catalog. This catalog contains nearly 100 million sources with visits that are at least 15 minutes long. The second step is to detect an important, complementary population of objects: those sources that did not have sufficiently high signal-to-noise to be included in the mission's source catalog, but do rise above the background noise during their brightening events.

There are several challenges in creating the Inter-Visit Variability Catalog. The first is computational: how do we construct the images and light curves needed for both the known sources and the inter-visit transients in a reasonable timescale? The second is algorithmic: how does one detect and classify transients in hundreds of millions of light curves and image frames, while minimizing false detections? One component will include a machine-learning approach to conduct an automated search, while the other will include citizen scientists engaged through a Zooniverse-powered user interface. The citizen scientists' dataset will provide a crucial training set for the automated algorithm, identify unique events that slip pass the automated classifier, and assist with completeness statistics.

A single server currently handles all gPhoton queries. To construct an all-sky catalog would require either dedicated access to a copy of the gPhoton database, or to distribute the database across multiple access points, so as to not impact all other users and to parallelize the process. It would also require a compute environment that would be able to create the light curves and data cubes in parallel, and to execute the automated searches for variability. While there are some machines available for computation at STScI, many of them are intended for use by *HST*-specific projects, and only have a few dozen CPU cores. ***This project would greatly benefit from a truly Institute-wide compute environment, with high-speed access to the gPhoton database, and with enough CPU cores to support gPhoton along with other demands from Institute researchers.***

Some steps to help enable the proposed *GALEX* Inter-Visit Variability Catalog would include storing a copy of the gPhoton database on a faster machine with solid-state drives for rapid access of the SQL queries, increasing the memory on the server to enable larger (and fewer) queries when building data cubes, adding more servers to help distribute the query load, and/or installing gPhoton on a cloud or high-performance compute environment (*e.g.,* XSEDE[9], Stampede at Univ. of Texas, Amazon). The latter would greatly speed-up the process of creating the data products needed to search for inter-visit variables and conducting the search itself.

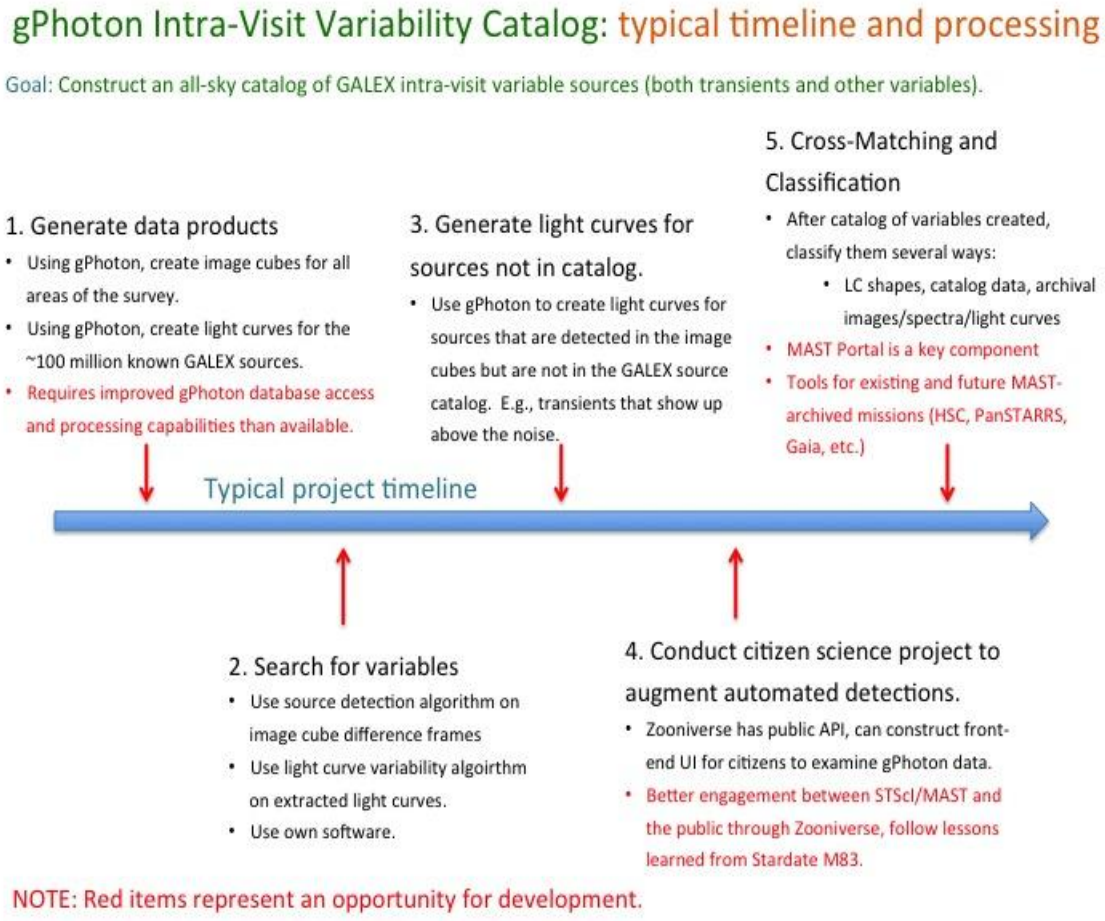

Figure 24: Workflow for the detection and characterization of transient phenomena in the GALEX gPhoton database.

To summarize, gPhoton will benefit from (1) a better database server or multiple database servers to ***increase the number of simultaneous queries it can handle*** and (2) a dedicated, high-performance compute environment ***to reduce the time needed to create data cubes and light curves, and to perform an automated search for inter-visit***

[9] XSEDE is a single virtual system for scientific computing. See www.xsede.org for more information.

variables. It is important for both of these improvements to proceed in parallel, since increasing only query capacity or computation capacity will bottleneck the other.

4.6 Multidimensional Exploration of Spectroscopic Datasets

Spectroscopy delivers a rich set of physical diagnostics to astronomers by measuring emission and absorption continuum and lines from a wide range of energy states and physical conditions. Spectroscopy allows astronomers to characterize ices in the densest star forming clouds, sniff molecules on exoplanets, and know how fast the first galaxies are forming stars. Because they arise in such a diverse set of astrophysical conditions, and add a third (wavelength) dimension to images, spectroscopic datasets present unique analysis opportunities and corresponding challenges.

The generic data analysis challenge in a spectroscopic dataset is the detection and characterization of features. These might be absorption lines from an intervening interstellar cloud detected in the spectrum of a background star, emission lines from a protostellar disk forming planets, or perturbations in a continuum and broad lines emitted from a super-massive black hole. These features may vary widely in their strength and location from source to source, even for similar objects, and may even vary with time in the spectrum of a single source. The essence of astrophysical spectroscopy is to robustly identify and measure these complex features and to extract physical insights from these measurements. The great diversity of astrophysical environments probed with spectroscopy guarantees that there is no one-size-fits-all solution for spectroscopic data analysis, though there are common approaches and pitfalls. This section considers those aspects of spectroscopic datasets that present particular challenges in terms of "Big Data", for two major science use cases.

4.6.1 Intergalactic and Circumgalactic Gas in UV/Optical Spectroscopy

UV spectroscopy with *Hubble's* four generations of spectrographs (FOS, GHRS, STIS, and COS) have taught astronomers almost everything they know about the distribution of diffuse gas in the Universe, from the intergalactic medium (IGM) between galaxies in its distinctive "cosmic web" to the circumgalactic medium (CGM) near galaxies that feeds their growth and later recycles their ejecta into new stars. We know that this diffuse gas contains a large share (40-60%) of all the normal (baryonic) matter in the Universe, and that its substantial enrichment with the heavy elements produced only in stellar interiors traces a vast cycling of matter in and out of galaxies multiple times over billions of years. These insights come from detections of absorption by IGM and CGM gas in the spectra of background quasars. These spectra are usually highly multiplexed in an unfamiliar sense: just as a deep-field image contains many galaxies or stars over a range of distances, a quasar sightline will usually intercept numerous unrelated gas structures along the line of sight in the IGM, the CGM of galaxies lying near the sightline, and the interstellar medium of our own Milky Way. The "absorption systems" arising at these different sites can overlap and

interfere with one another, each imprinting its unique pattern of lines at varying strength and complexity according to the physical conditions where they are produced. It is the diversity of these environments, translated into the complex patterns of lines they generate, that present the "Big Data" challenge arising from UV spectroscopy of the IGM.

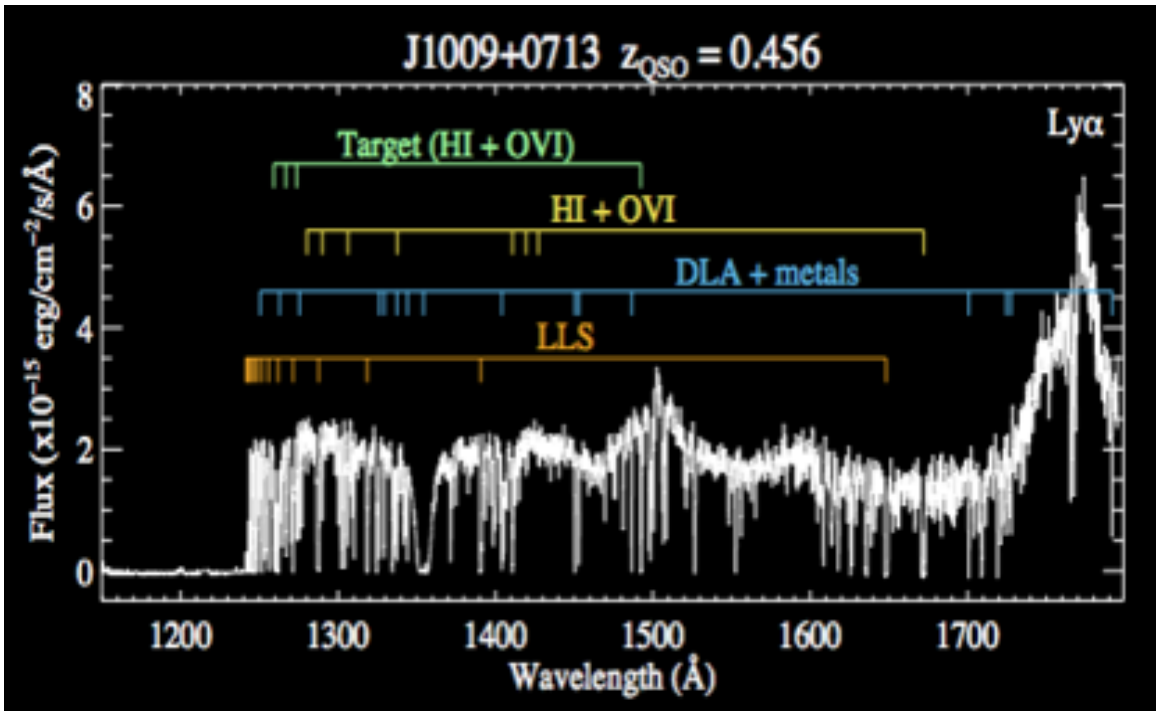

Figure 25: Co-added COS data, plotted versus observed wavelength and binned by three raw COS pixels to optimal sampling of two bins per resolution element. The Lyman Limit Systems (LLS) from Tumlinson et al. (2013) and the Damped Lyman-alpha Absorbers (DLA) from Meiring et al. (2011) are marked at the positions of Lyman series lines (orange), metal lines (blue), and O VI (green, yellow). The Lyman limit of the present system is clearly visible as the nearly complete drop in flux from the QSO spectrum below 1240 Å. Geocoronal Ly-α emission near 1216 Å has been excised.

Figure 25 shows a typical example of a QSO spectrum that exhibits a highly multiplexed and information rich probe of the IGM and CGM. This single COS spectrum obtained as part of the COS-Halos program (Tumlinson et al. 2013) was targeted at the CGM of the galaxy system labeled "Target (HI + O VI)". Because the sightline spans 1.8 Gigaparsec, it passes through many gas clouds along the way, each of which generates an "absorber system" with a unique pattern of lines (e.g., Lyman lines from HI, doublets from O VI and C IV in hot gas, and many "low ions" such as from cool gas containing Si, C, and Fe). These systems occupy single points in redshift space, but overlap in wavelength space because of the varying rest-frame wavelengths of the lines. The net result is a spectrum containing hundreds of diagnostic lines from at least 10 distinct absorber systems, all multiplexed into the same line of sight. And this is just one of the hundreds of such sightlines that *HST* has observed.

Thus, the "Big Data" problem for IGM/CGM spectroscopy consists of two key challenges:

1. Complexity: to disentangle these complicated datasets into their constituent systems by identifying, fitting, and cataloging the line IDs and measurements.

2. Correlations: to relate the properties of each system and its lines to the properties of all other such systems and to the measured properties of the nearby galaxies and the associated large-scale structures.

The full extraction of scientific knowledge from a spectroscopic database is primarily a "Data Discovery" challenge. For example, spectra from *HST* (and, eventually, *JWST*) are (will be) incredibly dense with diagnostic information about the temperature, density, elemental composition, kinematics, and fate of the gas as encoded in line strengths, ratios, profiles, and line-to-line correlations. Even more information is encoded in the relationships of all these properties to the broader properties of the nearby galaxies. And yet the identification and measurement of individual lines is still a process that is either largely manual or at least requires significant human intervention to make judgment calls about line IDs and the separation of blended lines. Automated identification of spectral features (e.g., the double absorption feature from Mg II in the UV) has been developed for some large homogeneous datasets such as SDSS (Zhu & Menard 2013). With the proper software tools, yet to be developed, the basic pattern matching approach that enables Mg II searches could be extended to more general patterns of lines and more complex spectra (e.g., those covering regions with many independent absorbers along a single sightline). ***This is the first challenge: to develop the tools of automated pattern recognition, machine learning, and statistical inference to enable automated or semi-automated processing of these rich datasets.***

Currently, correlations with galaxies hosting CGM or IGM absorption are generally done in "catalog space": line strengths and line widths are plotted against galaxy properties such as mass and SFR and patterns are sought and tested in the resulting scatter plots. But there are much richer possibilities that arise when one considers that each line is a non-ideal, even messy profile tracing fluctuating density and temperature, that each system imprints many lines on the spectrum, and that all this information usually is neglected because we lack the tools to use it properly. ***It is simply impossible to estimate, based on currently knowledge and limited exploration, how much untapped information already lies in Hubble's existing archive of UV spectroscopic data.*** Finding this out, and enabling our users to exploit this resource, is in STScI's direct interest as a consequence of our job to optimize the science return from the missions we support.

To exploit this potential, users need software tools for feature detection, correlation, and a data center facility for running these tools against the large datasets that *HST* has already generated. Development and release of such tools could begin at any time, as rich datasets already exist and the astronomical community interest exists to help develop and test, and then to use, these tools. ***Support for these tools in the same way that STScI has supported the similar source-detection efforts of the Hubble Source Catalog would greatly enhance the value of HST's existing archive for this important science***.

4.6.2 Galaxy Population Studies from Slitless Spectroscopy

Another form of spectroscopy that is big and getting bigger is slitless spectroscopy using grisms, gratings or prisms to take spectra of many objects simultaneously in a 2D format

such that each slitless spectrum retains spatial information about the targeted object. Slitless spectroscopy has become a growth area in *Hubble* observing, thanks to the powerful IR grisms on WFC3. It promises even more discoveries in the *JWST* and *WFIRST* eras, when the science instruments include some support for slitless and/or highly multiplexed spectroscopy. On *JWST,* one instrument (NIRISS) is highly optimized for this observing mode. Slitless spectroscopy presents challenges that are generic to most forms of spectroscopy - the identification and measurement of complex features - but also unique problems associated with overlapping spectra, highly background-limited image frames, and source extraction that hinders the community's use of slitless spectroscopy to its full potential. Figure 26 shows a simulated *WFIRST* slitless spectroscopic observation (only a small fraction of the actual data array is shown). This simulation is similar in quality to the data currently being obtained with the WFC3 grism on *HST.*

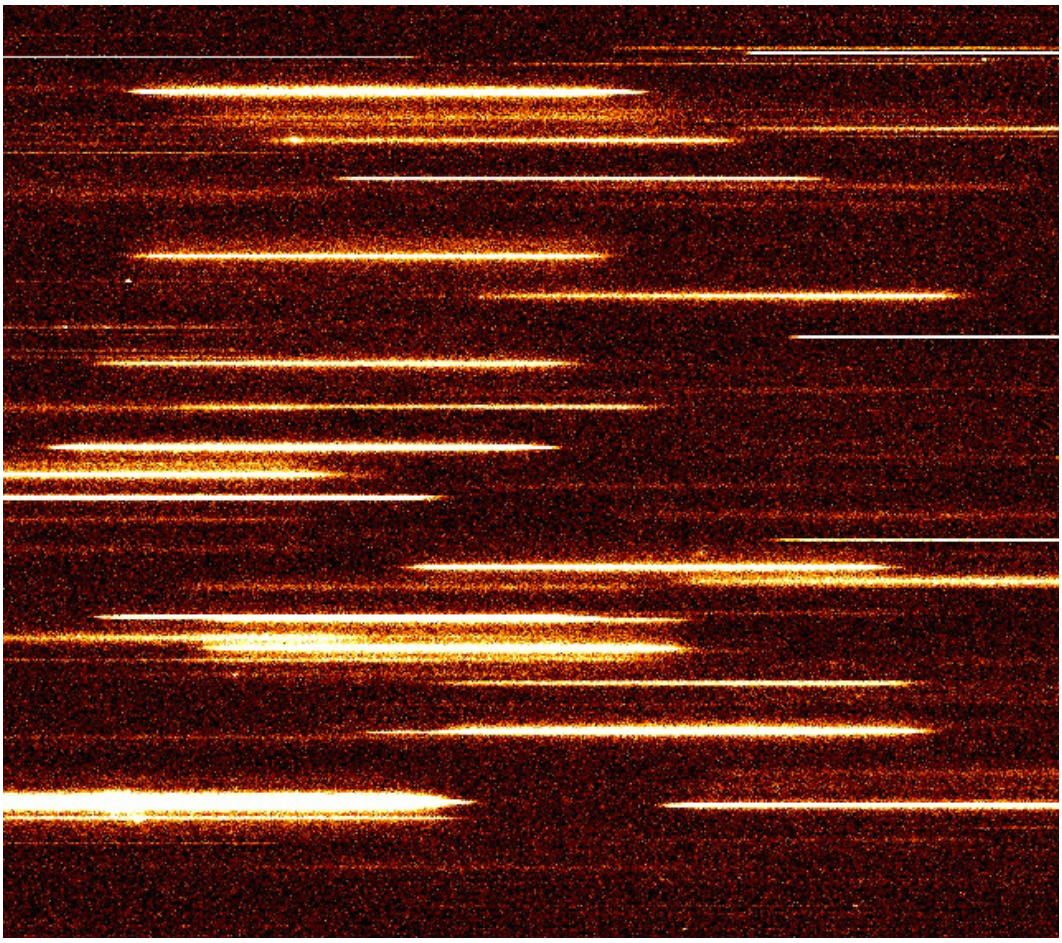

Figure 26: Simulated WFIRST slitless spectroscopic observation of a patch of sky. Most of the traces shown here are spectra of distant galaxies.

Hubble has already devoted hundreds of orbits to slitless spectroscopy, in particular with the 3DHST survey, the GLASS follow-up to the Frontier Fields, and some long-term pure parallel programs. So there is already a great reservoir of data to be exploited. With *JWST*, the flow will become a torrent, and this form of spectroscopy will present novel problems in

the management, reduction, analysis, and, ultimately, data discovery on large and complex datasets of slitless spectra.

Slitless spectroscopy is unlike the high-res UV spectroscopy discussed above in one key aspect that makes its challenges larger still. In UV IGM spectroscopy, it is rarely necessary to perform analysis or discovery directly on the original 2D detector frames because the spectrograph discards or blurs spatial information and the sources are typically points or point-like in any case. By contrast, slitless spectroscopy requires analysis of the 2D detector frames, and as the data gets deeper the problems of high backgrounds and more overlapping sources mount. These difficulties are then added to the source detection and feature identification (here, galaxy emission and absorption lines that diagnose their properties) to make for a complex and challenging problem in the extraction of insights.

Ultimately the user is trying to extract continuum shapes or line ratio diagnostics, and to perform correlation analysis on the resulting measurements. While some of this analysis can be done in catalog space, for some purposes it is likely that the user will want to test new reduction or extraction codes against the data frames, or to perform a forward model of the data based on physical models, in data space.

To exploit the full potential of slitless spectroscopy, users need software tools for spectral extraction, feature detection, correlation, and a data center facility for running these tools on the data hosted at that center. As for the 1D spectroscopic tools above, support for these tools at a level similar to that devoted to the HSC, for example, will greatly increase the value of *HST*'s archive for galaxy formation and stellar evolution science.

4.6.2.1 The Big Data challenge: Complex Spectroscopic Data

Spectroscopy is generally a "Big Data" problem by virtue of the complexity and dimensionality of the datasets, not their volume. For scale, there are about 13,000 1D-extracted spectra in the public COS archive and about 30,000 in the public STIS archive. The 13,000 COS exposures occupy about 30 GB on disk - not a "Big Data" problem in the classic sense. However, this dataset is extremely complex and structured even when restricted to the approximately 600 quasars that have been observed for absorption line measurements of the IGM and CGM. In this corpus of data, about 40% of the public COS archive, there are about 5,000 individual exposures and thousands of scientifically interesting absorption lines. The STIS archive adds to this with hundreds more targets and thousands more lines. Part of the complexity of this dataset arises from its heterogeneity, as the targets have been observed by two or more instruments, multiple gratings and settings, and may vary over time when observed more than once. While the usual end-user tasks of data combination and basic analysis are still workstation scale problems, the issues arising from feature detection and cataloging are extremely complex and will require concentrated development effort to bring to fruition. Spectroscopic datasets present a "Big Data" problem in the sense of "We have a lot of data and don't know quite what is in it". The science upside is significant if STScI can make progress in the solution of this "Complex Data" problem.

The limits of current tools: While numerous homebrew tools for line identification and analysis have been developed by collaboration of users, and there are some line-extraction algorithms that apply in limited circumstances (finding strong Mg II doublets in SDSS spectra), there is no community standard code that works in general cases. That is, ***there is no SExtractor tool for spectroscopy. This is the first step to tackle.***

Computing requirements: Mining of current and near-future spectroscopic datasets, as described, does not present challenges to modern computing infrastructure. ***The challenge is all in developing smart software*** as described above.

The Future Timeline: Mining of spectroscopic datasets is a key area of *HST* science where relatively modest investments could yield significant dividends immediately. As noted above, there is already a large archive of public data from *HST's* spectrographs. STScI has deep expertise in the analysis of spectroscopic datasets. *HST's* Spectroscopic Legacy Working Group has brought post-pipeline spectral combination to the point where most of the COS archive will be science-ready in 2016. This will soon be applied to STIS data as well.

In the long term, these tools could be adapted and tuned to encompass 2D spectroscopy, to mine the *HST* grisms archive, and then eventually to the single object, IFU, and MOS modes of *JWST*, and eventually *WFIRST*. STScI should undertake the initiative to develop automated spectral feature identification and classification software for use on 2D slitless spectroscopy.

5 Capabilities, Tools, and Science Drivers

5.1 STScI's Current Capabilities

5.1.1 Hardware / Computing Architecture

Up until approximately mid-2015, the computing infrastructure at STScI was developed with the "one server / one purpose" systems engineering philosophy. Little of the environment was virtualized, meaning that computing resources available to science staff are generally fixed and inflexible.

In addition to individual science workstations, STScI has 5 science servers available for research purposes. These servers range from 8 to 32 cores each and 32 to 400 GB of memory; all are running HTCondor[10]. Individual utilization of the servers is managed via informal communications and the honor system; jobs are submitted at the individual scientist's discretion. This may result in conflicts between research efforts with long running or CPU intensive workloads. The science servers are typically used for workloads such as image processing (e.g., astrodrizzling of ACS/WFC datasets), simulations and development and testing of software for various applications. Datasets range from a few gigabytes up to terabytes.

The physical nature (i.e., limited overall capacity, non-virtualized) of the science servers means that in addition to utilization conflicts, the compute environment cannot be easily reconfigured to handle different configurations as appropriate or optimal to the research being conducted. The clusters are frequently at maximum utilization and cannot be reconfigured to address peaks in demand. New servers are often "consumed" almost immediately.

Understanding the STScI functional compute environment is also important as excess compute or storage capacity in mission systems can be made available for research while STScI transitions to a virtual-first approach in server architecture. In short, STScI's computational infrastructure needs to be highly virtualized across the board to meet many (but not all, as described below) of the computational needs associated with a wide variety of science programs. This can be done in a way that ensures mission systems are not affected, limited, or otherwise put at risk.

HST mission systems at STScI are highly physical and fixed capacity, though the use of High Throughput Computing (HTC) via HTCondor does provide job scheduling and resource allocation capabilities for the DMS. *HST* mission systems are a mix of Intel and Sun

[10] HTCondor is a workload management system for compute-intensive jobs. It provides management of job scheduling, job prioritization, and resource monitoring. HTCondor is a free, open-source software package.

Microsystems (now owned by Oracle) hardware and a mix of operating systems such as Windows, Solaris, and Linux. Virtualization of *HST* servers is being reviewed on a case-by-case basis as part of the standard hardware refresh process. *JWST* mission systems are being virtualized from the ground up as new systems come online. Some in-house applications (such as DMS) that are primarily CPU-bound may stay physical, as this is still optimal in limited cases.

STScI is working on the implementation of a brand-new, production-class, fully-virtualized compute environment termed the "Flexible Data Center" (FDC). The FDC design was built from the ground up in partnership with VMWare and based on ESX 5.5 and NSX/Software Defined Networking (SDN). Networking has been designed to optimize inter-FDC data flows. STScI will be migrating servers into the FDC on an incremental basis. This process will take several years to implement.

Even with the evolution of the FDC, it is unlikely that any previously unused capacity gained in the physical to virtual transition will satiate the needs for big data analyses within the physical STScI data center. As of late 2015, STScI had an aggregate of about 2,000 cores as a theoretical maximum assuming a fully virtualized environment. In reality, it's likely that the number of cores that can be used for science will remain in the mid 100's. This will not meet the needs for Big Data analyses like those discussed in Chapter 4.

5.1.2 Network

The STScI data center is a mix of 1 and 10 Gbps networks. Typically, upgrades to 10 Gbps are made during normal refresh cycles or earlier to meet specific engineering requirements (e.g., connection to storage). Early upgrades to 10 Gbps will be made, as needed, to align with the increases in Internet/2 bandwidth. Internally, the Flexible Data Center will be 10 Gbps capable from the ground up.

STScI has a 500 Mbps connection to the Internet and a 1 Gbps connection to Internet2. While the existing links are sufficient for existing traffic (Internet usage is comparatively very low), plans are being made to increase the Internet link to 1 Gbps in the short term with 10 Gbps upgrades planned for both links when additional network upgrades are completed. In addition to preparing for *JWST*, these increases will be needed to support new large-scale archives such as PanSTARRS and *TESS*.

5.1.3 Storage

STScI houses about 8 Petabytes of storage using a typical mix of NAS and SAN approaches. The largest storage utilization is found with PanSTARRS (1.8 PB), followed by MAST/HLA (0.8 PB). Under current plans for *HST*, *Kepler*, *TESS*, and *JWST*, the STScI mission archive will grow to 3.0 PB by 2018 and to 5 PB by 2021 (see **Figure 2**).

5.2 Overview of Needed Capabilities

As noted, the existing STScI compute environment is oversubscribed for even "traditional" use cases. The science cases outlined in this report will require new approaches and dedicated funding to convert this deficit into a state-of-the-art environment capable of supporting the needs of our users and our missions in the next few years.

Table 3 summarizes the high-level requirements for hardware and software derived from the science cases presented in Chapter 4. While the numbers of CPU cores cited for each science case are not highly precise, the estimates run from 500 to over 10,000. Estimates of near future storage needs are also significant and indeed top off at around 10 PB. If we include the need to store relevant simulated data, the storage requirements could easily grow to 20 to 30 PB. Bandwidth requirements are between 1 Gbps and 10 Gbps. The greatest gap between current STScI capabilities and the science use cases presented here is the availability of high performance computing (number of CPU cores). Figure 27 presents a summary of the science case computing requirements in graphical form, along with our estimate of our current science computing capabilities.

While additional storage is also needed, it is the easiest component of our infrastructure to expand. Advances in data storage (size, price, etc.) have actually outpaced the trend in CPU power most commonly identified as "Moore's Law" (which, after holding true for many decades, is starting to flatten). STScI has been, and will continue to be, able to keep up with data growth. The challenge is in our ability to find meaning and value from these data. Bandwidth growth, in contrast, significantly lags behind both storage and CPU trends (and this is likely to continue to be the case); since analyses require the constant flow of data from/to compute and storage, the main issue lies in mitigating the data "nearness" problem. This trend requires new approaches and capabilities.

Network resources are often the bottleneck in any approach, but there are not as many options available to STScI as it is difficult to scale bandwidth in the same way as one scales CPUs or disks. The bandwidth "between" the CPUs/memory and the data is what is most critical; there is a big difference between local data center bandwidth and Internet/Internet2 bandwidth needed for use of non-local data sources.

When looking at bandwidth, there are two key points to consider. First is the speed at which the CPU/memory can communicate with local SAN, or NAS-based storage. There are many mature technologies that meet this need. At present, this is not the primary constraint for STScI and those in the community wishing to make use of the STScI-housed archive data using traditional approaches. Second, and more relevant to this discussion, is the bandwidth between the CPU/memory and the data when the data are not located at STScI (or conversely, the CPU/memory is remote while the data are local to the STScI data center). This is where Internet/2 bandwidth is crucial – either for direct data analysis or for transfer of the data to be "near" the compute resources. In reality, using Internet/2 at the 1

to 10 Gbps level is not feasible for large data analysis when approaching the Petabyte range or even simultaneous multi-Terabyte data transfers.

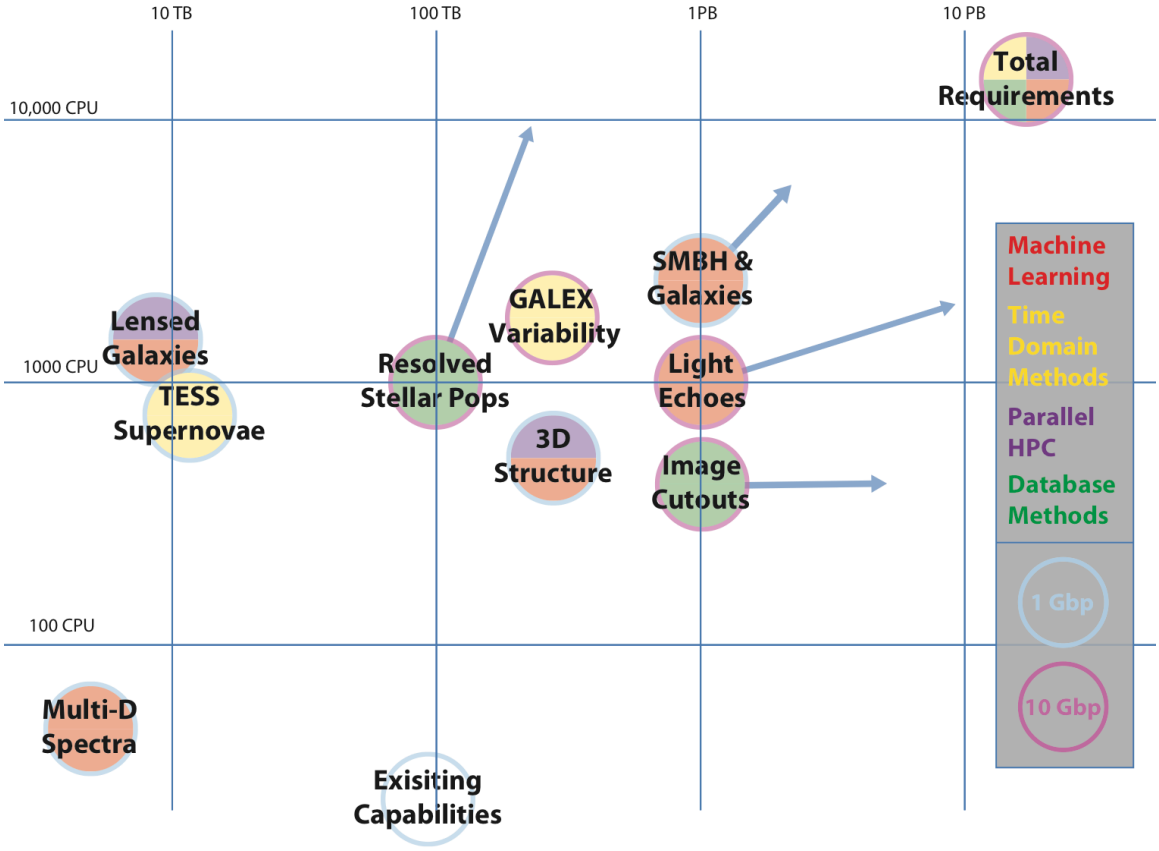

Figure 27: A visual representation of the computational (hardware, bandwidth, services) and methodological (software) needs of each "big data" science program presented in Chapter 4. The X-axis is storage needed; the Y-axis is size of a cluster needed, as measured in the number of CPU cores. The filled colors of the circles represent a short hand for kind of software methods needed: machine learning (red), time-domain methods (yellow), parallel HPC (purple), and database methods (green). The circle color represents the bandwidth needs — 1 Gbps in blue, 10 Gbps in pink. Arrows represent the range toward more required capabilities for more advanced versions or expansions of the projects. The Institute's existing science computing capabilities (bottom center) and the combined requirements needed to support all the big data science use cases simultaneously (top right) are also shown

While STScI is planning to proactively increase bandwidth as needed, this does not fundamentally address the nearness problem. Instead, ***we have to look at new architectures that enable the astronomer's compute environment to be "close" to the large datasets being analyzed.*** This requires more effective use of HTC/HPC as well as virtual desktop infrastructures (VDI). Because setting up these capabilities is complex, standard practice is to do so in a fully virtualized environment; the alternative of reconfiguring existing physical environments is slow and inflexible. While it is not unheard of to have a physical environment designed optimally for the purpose of big data analyses, it is not suggested as the primary option because of cost and lack of flexibility.

New computing capabilities, however, require new approaches to the tools and programming methods used by astronomers. The growth in CPU performance has largely been the result of transition from single core to multi-core to many cores with hyper-threading technology. Processing capability is no longer defined just by processor GHz, but by the number of cores available to the programmer / astronomer. A single workstation can enable the astronomer to take advantage of multiple cores and analyze a few hundred or few thousand images, but such tasks can quickly reach and exceed the computational capacity of a desktop machine. Instead, it is necessary to move to a "many" core approach where the number of available cores can be in the hundreds to thousands range.

In addition to increased use of parallelization, new approaches to store and retrieve data (with a focus on using commodity hardware) have emerged as essential when working with "big" data. Hadoop is the most commonly known approach for managing distributed storage and data processing of these very large datasets. In short, while STScI uses HTCondor for distributing computation workloads, ***more effort and focus should be placed on implementing advances in data storage and data management.***

Table 3: Computing and Software Capabilities Needed

Science Case	Computing	Storage	Bandwidth	Software
Light echo detection and classification	>1000 CPU core cluster	~1 to 10 PB	~10 Gbps needed for citizen science and data transfer	Machine Learning (ML) classification, feature vectors, citizen science user interface.
Lensed galaxy detection	~1000 CPU core cluster	~10 TB (if image cutout service is available for large surveys).	~1 Gbps	Parallelizable code management; ML algorithms.
Resolved stellar populations	>1000 CPU core cluster (~10K core ultimately needed)	few 100 TB	~10 Gbps (uses non-local data)	Automated pipeline management s/w. Efficient database query tools.
3D Structure in the cosmos	~500 core cluster	>200 TB	>1 Gbps	Parallelizable code management; clustering algorithms; photo-z algorithms.
SMBH / Galaxy Co-evolution	>2000 core CPU	>1 PB	>1 Gbps	ML; feature vectors, efficient cross correlation algorithms; high dimension data visualization tools
TESS supernova search	~1000 CPU core	~10 TB for raw data	~1 Gbps (if data local)	Transient detection algorithms for highly undersampled data
GALEX Variability Catalog	High performance computing required. Exact specs TBD.	few hundred TB, solid state disks for fast access.	High bandwidth to local db (~10 Gbps)	Algorithms to detect and classify transients from archive of 10^8 light curves.
Multidimensional spectroscopic datasets	Current probably OK.	~100 GB.	Current OK	Feature vectors, ML algorithms. Automated spectral feature detection software. Efficient cross-correlation algorithms.
Image cutout service for wide-area sky surveys	Dedicated high-user volume server needed.	>1 PB	10 Gbps or more	Must support thousands of users and many simultaneous queries.

5.2.1 Data Integrated Visualization

To be able to make discoveries with Big Data hosted at STScI we must enable scientists, both local and global, to explore our data. This has been very difficult in the past; many important novel objects and trends are still being discovered in SDSS because exploratory access to the dataset is cumbersome. Large or complex queries are slow, then the data must be downloaded slowly, and then custom visualization tools for inspecting these datasets must be employed, as the data are high dimensional and often too large for memory.

STScI can circumvent this problem by building powerful, flexible, and integrated data visualization tools that sit server-side, allowing users to quickly generate scientific hypotheses that can then be more rigorously tested in the science cloud or by downloading the appropriate subset of the dataset. This would be a tool broader than what is currently available in MAST, but narrower than the fully flexible environment we would provide in a science cloud.

One working example we have already invested in is the *Glue* architecture. *Glue* allows users to explore data through linked, brushed views of histograms, images, and scatter plots, and connect multiple datasets across wavelength for simultaneous manipulation. *Glue* is a "hackable interface"; it is written in Python for easy direct manipulation by astronomers, and designed modularly to allow astronomers to add their own custom importing, viewing, and manipulation tools. At present *Glue* works only on desktop clients and with relatively small datasets, but could be adapted to a web-interface for large datasets.

Machine learning and deep learning, described in Section 4.1.1, also provide an important tool kit for data-integrated visualization. Much of visualization is about grouping data for more efficient selection. Machine learning methods like t-Distributed Stochastic Neighbor Embedding (*a.k.a.* t-SNE; van der Maaten 2014) and deep learning methods like auto-encoders are capable of taking very rich datasets and compressing them down to two dimensions for easier visualization and exploration. ***The boundary between data exploration and dimensionality reduction will continue to blur in the big data era***.

5.3 Needed Capabilities in the Next 2 Years (Pre-*JWST* Era)

In the next two years, STScI will continue to reconfigure its computing environment, following the recommendations not only from the Big Data Science Definition Team, but also from an internal working group comprised of IT, mission, science, and engineering staff. These inputs will define how the STScI Flexible Data Center will evolve to meet our collective needs. Note that as stated previously, we don't expect to net more than a few hundreds of cores (assuming no net investment for this specific purpose). While this is a big step forward from the existing science server clusters available to STScI staff today, it is still a far cry from the 1,000's+ needed for big data analyses.

In the next two years, ***STScI should explore options for many-core HPC outside of the existing science servers***—bigger, faster desktops are not viable for many of the big data questions we seek to solve. Instead, this would be done onsite via the Flexible Data Center (FDC) and the implementation of "private" cloud capabilities, offsite via partner institution compute resources, or offsite via public cloud services as offered by Amazon or Microsoft. Use of public cloud/other institution is of course constrained by available bandwidth (and the "data proximity" problem), but will be the only way to get to the 1,000+ CPU core level.

Another risk to the implementation of these new science compute capabilities is if cost constraints require STScI to downsize computational capabilities by attempting to maximize efficiency in the functional environment. In other words, the entire premise of enabling science via the Flexible Data Center hinges on the availability of unused CPU and memory in the functional environment. As the functional environment makes more and more efficient use of computing resources via virtualization, this could mean that there is less "unused capacity" left that can be made available for science.

Regardless, most of the big data use cases identified in this report will not be feasible in the short to medium term STScI-compute environment. ***A pilot project that leverages a public cloud commercial service and/or other academic institution should be initiated as soon as possible as it is not dependent on the FDC***. This will also require mirroring key Archive data to the hosted compute environment. In addition to an effort that leverages public cloud availability of cores on demand, a pilot effort to build more expertise in alternative data management (e.g., Hadoop) should also be considered.

5.4 Needed Capabilities in 3 to 5 years (*TESS*, *JWST* Era)

The biggest long-term challenge for STScI is the "nearness" problem. This will manifest itself in several ways:

- An astronomer outside of STScI wants to use big data analytic techniques with the STScI archives.
- An STScI astronomer wants to leverage large-scale dataset at another institution.
- An astronomer (STScI or other) has access to a third-party (other educational institution or public cloud provider) computational platform, but the data exists elsewhere.
- An astronomer (STScI or other) wants to cross-reference multiple, large-scale data physically located at different institutions.

These challenges are addressed today largely by the use of data mirroring. This puts a copy of the data "near" the scientist's workspace. However, as stated previously, bandwidth is not even remotely expected to keep pace with the growth in data volumes. New approaches are needed.

In the simplest of terms, the scientist needs:

- Data to analyze;
- A workspace with software and tools to analyze the data (e.g. a desktop);
- Computing resources (CPUs, memory, storage);
- Bandwidth connecting the data, workspace, and computing resources.

If it is increasingly difficult to move the data due to bandwidth limitations, then it will be necessary to instead move "the scientist." This can be accomplished by creating the capacity to co-locate the data and computing resources and in some cases the workspace—preferably at the LAN level, but also by the addition of very high speed external network connections (100 Gbps+ links, may be cost-prohibitive for STScI). The scientist can remotely connect to the workspace from anywhere on Earth.

Co-locating the workspace with the data will require STScI to create a virtual desktop infrastructure (VDI). VDI will enable STScI to put the astronomer's workspace close to the compute resources and the data. While this is not a problem for STScI staff working in STScI buildings, it is a challenge when astronomers from the community want to take advantage of STScI's image archives on the big data scale. This also assumes, of course, that STScI provides computing resources onsite for the remote astronomer, which could be possible on a small scale, but infeasible as datasets grow in size.

STScI needs to expand its definition of the "data center." Serious consideration should be given to replicating large STScI-generated datasets and data products to external third parties with scalable compute infrastructures. STScI could also provide VDI capabilities at that same third-party location.

In summary, in the next 5 years, ***STScI will need to invest in a hybrid cloud (FDC/private + public cloud/partner institution), even larger high-speed external data links, and a virtual desktop infrastructure***. Unless STScI has both high-speed data links and *local* computational capability, the "big data" value of STScI's image data will be minimized.

5.5 Capabilities Beyond 5 years (*JWST*, Pre-*WFIRST* Era)

More than 5 years out, STScI will need to establish partnerships with other organizations and enhance the "mobility" of data, compute, and workspaces. This means that in addition to very high speed links to other organizations, STScI will need to become part of a larger multi-organization compute environment that facilitates the transfer of data, access to compute resources and workspaces to best optimize our capabilities. This is the ultimate expression of hybrid (public + private) cloud approaches. This is not yet being done on a wide scale.

If STScI cannot offer sufficient in-house computational capability, then it will have to partner with another organization or leverage on-demand services such as the public cloud. This requires solving or reducing the "nearness" problem through some other means.

STScI may want to consider, possibly for *WFIRST*, a "cloud first" option where the data storage and compute environment are designed to be housed at a third party / public cloud where bandwidth is less of an issue and cores/storage can be readily scaled up (and equally valuably scaled *down*) with relative ease. In addition to the technical changes needed, the funding model would need to adapt as well.

6 Skill Sets for Big Data Science

Big data science will require STScI to build up its skill base in a number of areas. We highlight the ones we feel are most important and urgent in the near term (next 1 to 5 years).

6.1 Data Science Postdoctoral Fellowship

The Institute should ***initiate a new postdoctoral fellowship program that is focused specifically on hiring an early-career researcher whose interests and expertise are centered on big data science***. The fellowship should be designed to provide up to 3 years of support for an outstanding postdoctoral researcher, working here at STScI on innovative scientific studies involving large astronomical databases, massive data processing, data visualization and discovery, automated classification and/or clustering algorithms. Emphasis would be placed on researchers whose studies take significant advantage of the archived datasets held (or planned for ingest) at STScI/JHU, including *HST, GALEX, Kepler/K2*, PanSTARRS, *TESS, JWST*, and *WFIRST*. The fellowship holder should be encouraged to consider interdisciplinary approaches to addressing challenges posed by big data in astronomy, including facilitating such studies by a broad user community, and exploit the potential for partnering with JHU initiatives. ***It would be ideal to identify the funding to be able to support at least two simultaneous data science postdoctoral fellows***. This will allow the fellows to collaborate with one another as well as with the full time STScI staff.

6.2 Archive User Support Skills

Learning how to use new software can be frustrating and time-consuming, and can be a limiting factor in a person's willingness to even try the tool. Effective user support can be critical. While videos and graphic use-cases can be helpful, it is the willingness of a person to provide a little timely advice that often gets a person over the initial hurdle.

For this reason, ***STScI should plan to provide adequate user support for advanced data visualization and machine-learning tools that are developed for big data programs***. To date, our user support has included the STScI Helpdesk and the contact scientist program (for both observing proposals and archival research projects). To be successful in the future, the products and tools developed as part of our big data initiatives will also require effective documentation and user support. While much of the infrastructure (e.g., help desk software) and resources (including human) already exist, they may require "modernization", including staff that can provide advice in using the new generation of knowledge-based algorithms and tools.

An additional component of this enhanced level of user support could include working more closely with outside investigators, for example exercising the same workflow that they are using, to help diagnose possible obstacles that they currently face in trying to use our

software and our archive interfaces to achieve their science. This would have broader benefit, by providing us with very direct feedback on current limiting factors, and enabling us to find new ways to improve these systems to make them as useful as possible for the community as a whole. Specifically, this could include user support for server-side APIs and scripting environments as well as staff who are comfortable using machine-learning techniques.

6.3 Skills to Develop Data Analysis Tools and Science Products

STScI should invest additional effort on producing high-level science products and tools to enhance archival science. Existing examples of high-level science products for *HST* imaging data are the *Hubble* Legacy Archive (HLA) and the *Hubble* Source Catalog (HSC). A similar level of activity should be aimed at developing a software equivalent of "Astrodrizzle" for co-adding spectroscopic datasets and a software equivalent of "SExtractor" to identify, deblend, and measure absorption/emission lines in spectra. Generic tools for visualizing and analyzing time-domain astronomy data should also be created and developed.

6.4 Organizational Structure within STScI

The organizational structure adopted to support Big Data science within STScI will play an important role in achieving the ultimate goals of this project. STScI currently supports or is developing a range of archival activities, including the holdings under MAST, the PanSTARRs database and the nascent *JWST* archive. ***Consolidating the leadership and oversight of those activities will lead to greater coordination, eliminate duplication, enable more efficient use of resources and allow the Institute to take full advantage of synergies between the data science work on our multiple missions***. Prior discussions have centered on the concept of a Data Science Mission Office (DSMO), with overarching responsibility for supporting all archival holdings at STScI. ***We recommend that STScI form a Data Science Mission Office as soon as possible***. The prime focus of the DSMO should be maximizing the scientific potential of the multi-wavelength archival data; thus, the focus goes beyond simply enabling easy data access to providing the community with tools that can exploit the full potential of multiple disparate datasets. To realize this potential, DSMO must bring together four critical areas of STScI excellence in current and future staff: software engineering, statistical and mathematical methods, astronomy, and information technology.

A Data Science Mission Office would include at least three key personnel:

- The Data Science Mission Head, a Directorate-level position, will be responsible for setting the overall strategic vision for the archive science at STScI. The mission head will coordinate and manage all archive activities at STScI. The mission head will be an advocate for the community's scientific stake in developing interdisciplinary approaches to addressing the challenges posed by big data in astronomy. The mission head will also be the primary STScI interface to the broader astronomical data sciences community, working in concert with colleagues within STScI and

NASA to promote archival science to various committees, panels and decision makers.

- The DSMO Project Scientist will be responsible for developing and maximizing the scientific impact of the Data Sciences Mission. They will provide leadership in identifying, defining and conducting science-driven trades and improvements to mission services, products and processes. The Project Scientist will take a lead role in working with the astronomy community, both internal STScI staff and our external user community, to develop long-range scientific initiatives for Archives at STScI.

- The DSMO Project Technologist will be responsible for developing and utilizing appropriate technologies to support archive science. They will track trends in the broader community, and work with STScI engineering and information technology to set appropriate priorities for hardware and software development.

These three personnel will provide the core leadership of the Data Science Mission; staff members from other STScI missions and division will be matrixed with DSMO to support its activities. In particular, a primary objective of the DSMO will be to foster synergies between the Institute's archive science branch, science software branch, and science instrument teams to create, develop and deploy high-level science products and the tools needed to visualize and optimize the extraction of astronomical information from the complex and/or large datasets we host now or that we will host over the next decade.

For DSMO to succeed in enhancing the effectiveness and efficiency of our archive expertise, we believe it will be essential that the DSMO Head be given the authority to prioritize spending on archive initiatives. This may entail prioritizing spending that transcends mission budget boundaries when initiatives are identified that are needed by, or would provide high science-added value to, those missions. The DSMO Head will work in conjunction with the other mission offices to ensure that all our critical archive requirements are met while still providing ample opportunity to innovate for the future of astronomical data science.

7 Summary of Recommendations

The SDT has considered a range of plausible science use cases that involve what we would define as "big data" astrophysics. These science cases all require efficient access to STScI's present and near future data archive holdings as well as the tools to visualize and extract their information content. Based on our assessment of the changing nature of astronomical research and the level of our current computing infrastructure we believe that the following recommendations must be seriously considered to ensure that STScI remains a world-class data archive research facility. We have grouped the recommendations into three main categories: Infrastructure, Software and Tools, and Organization and Workforce.

7.1 Computing Infrastructure

Infrastructure recommendations are focused on enhancements to our networks, storage systems, and science computing capability. This report identifies the following key infrastructure recommendations:

1. External Network bandwidth: Upgrade the bandwidth on all external networks to 1 Gbps by end of 2016. Upgrade the bandwidth on all external networks to 10 Gbps by the end of 2018.
2. Internal Network bandwidth: Upgrade the bandwidth on all internal networks to 10 Gbps by 2018. Explore options for upgrades to 100 Gbps internal bandwidth by 2021.
3. Storage Capacity: Increase data storage in MAST to 6 PB by 2018. Increase storage capacity to 30 – 60 PB by 2021.
4. Computing Power: Obtain access to multiple virtualized, dynamically configurable nodes with upwards of 1000 CPU cores each by the end of 2018. Access to 10K core systems should be sought by 2021. These goals may be achieved via a hybrid approach that utilizes both the improved computing environment being built into the STScI Flexible Data Center and also utilizes partnerships with external HPC facilities (e.g., JHU's access to the Mid-Atlantic Regional Computing Center). Given the significant head start many external HPC centers have, a joint venture seems to be the most cost effective approach to achieving access to the needed computational power that will be required for astronomy in the next few years.
5. Increase database server capacity to allow high-volume query usage. Two systems that should be given priority are the servers associated with MAST's gPhoton service and the servers that handle our image cutout services.

7.2 Software and Tools

The demands of astronomical data discovery and analyses in the coming years lead us to make the following recommendations for our science software environment:

1. Build a user environment that allows users to upload scripts and run API's on our systems and in close proximity to our large data storage facilities. This requires more effective use of HTC/HPC as well as virtual desktop infrastructures.
2. Support the development of integrated data visualization tools that sit server-side to enable researchers to efficiently explore all our data holdings.
3. STScI should explore, select and deploy machine-learning architectures to support archive researchers who will, more frequently, require advanced classification and regression analyses to extract scientific results from our data holdings.
4. STScI should initiate the development of automated spectral feature extraction and classification software. The software should be adaptable to data from all spectrographs on current and future missions being hosted in MAST.

7.3 Organization and Workforce Skill Mix

There were several additions and changes to our organization and our workforce skill mix that we feel are needed to allow the Institute to provide the most effective world-class science archive facilities in the big data era.

1. Initiate a Data Science Postdoctoral Fellowship program focused on early-career researchers whose interests and expertise are centered on big data science. Emphasis should be placed on researchers who can help augment our big data capabilities in one or more areas over the course of their 3-year tenure at STScI. Ideally, this should be an ongoing, long-term postdoctoral fellowship opportunity that will be viewed as on par with other prestigious fellowship programs. Identifying the resources to support at least two data science fellowships per year should be an important goal for the Institute.
2. Expand our user support team to include staff with expertise in server-side APIs and scripting environments.
3. Expand our staff expertise in the areas of machine learning and automated classification techniques.
4. Expand our staff expertise in the development and application of code to reduce the dimensionality of highly complex datasets.
5. Consolidate the leadership and oversight of all of our data archive activities to a single Data Science Mission Office. The head of the DSMO should be a Directorate level position. The DSMO would also host a project scientist and a project technologist. The DSMO Head should be given the authority to prioritize mission archive-related funding across our full mission portfolio while ensuring all key archive milestones and requirements are met.

8 References

Badenes, C., et al. 2008, ApJ, **680**, 1149.

Badenes, C., et al. 2009, ApJ, **700**, 727.

Baldi, R. D., Chiaberge, M., Capetti, A., et al. 2013, ApJ, **762**, 30.

Beaumont, C. N., et al. 2014, ApJS, **214**, 3.

Beck, R., Dobos, L., Yip, C.W., et al. 2016, MNRAS, **457**, 362.

Bellini, A., et al. 2010, AJ, **140**, 631.

Benitez, N., et al. 2000, ApJ, **536**, 571.

Bertin, E., & Arnouts, S. 2010, Astrophysics Source Code Library, ascl:1010.064

Bianco, F. B., et al, 2011, ApJ, **741**, 20.

Bloom, J. S., et al. 2012, PASP, **124**, 1175.

Boldrin, M. et al. 2016, MNRAS, **457,** 2738.

Bolzonella, M., Miralles, J.M., & Pelló, R., 2000, A&A, **363**, 476.

Botzler, C. S., Snigula, J., Bender, R., Hopp, U., 2004, MNRAS, **349**, 425.

Brammer, G. B., van Dokkum, P. G., Coppi, P., 2008, ApJ, **686**, 1503.

Brink, H., et al. 2013, MNRAS, **435**, 1047.

Calzetti, D., et al., 2015, AJ, **149**, 51.

Coe, D., Benítez, N., Sánchez, S. F., et al. 2006, AJ, **132**, 926.

Dalcanton, J., et al. 2009, ApJS, **183**, 67.

Dalcanton, J., et al. 2012, ApJ, **200**, 18.

DeGraf, C., Di Matteo, T., Treu, T. et al. 2015, MNRAS, **454**, 913.

Di Matteo, T., Springel, V.T., & Hernquist, L. 2005, Nature, **433**, 604.

Dieleman, S., Willett, K.W., Dambre J. 2015, MNRAS, **450**,1441.

Ferrarese, L. & Merritt, D. 2000, ApJ, **539**, L9.

Ganeshalingam, M., Li, W. & Filippenko, A.V., 2011, MNRAS, **416**, 2607.

Gebhardt, K., Bender, R., Bower, G., et al. 2000, ApJ, **539**, L13.

Grogin, N. A., Kocevski, D. D., Faber, S. M., et al. 2011, ApJS, **197**, 35.

Gultekin, K., Richstone, D., Gebhardt, K., et al. 2009, ApJ, **698**, 198.

Hayden, B., et al. 2010a, ApJ, **712**, 350.

Hayden, B., et al. 2010b, ApJ, **722**, 1691.

Heckman, T. M. & Best, P. N. 2014, ARA&A **52**, 589.

Ho, L. C. 2008, ARA&A, **46**, 475.

Hopkins, P., Murray, N., Thompson, T. A. 2009, MNRAS, **398**, 303.

Horesh, A., Ofek, E.O., Maoz, D., et al. 2005, ApJ, **633**, 768.

Huchra, J. P. & Geller, M. J., ApJ, **257**, 423.

Huertas-Company, M., Gravet, R., Caberra-Vives, G., et al. 2015, ApJS, **221**, 8.

Kasen, D. 2010, ApJ, **708**, 1025.

Koekemoer, A. M., Faber, S. M., Ferguson, H. C., et al. 2011, ApJS, **197**, 36.

Krause, O., et al. 2008a, Science, **320**, 1195.

Krause, O., et al. 2008b, Nature, **456**, 617.

Marriage, T. A., Acquaviva, V. A., Peter A. R., et al., 2011, ApJ, **737**, 61.

Meiring, J., et al. 2011, ApJ, **732**, 35.

Olling, R. P., et al. 2015, Nature, **521**, 332.

Piro, A. L., & Nakar, E. 2013, ApJ, **769**, 67.

Postman, M., Coe, D., Benitez, N., et al. 2012, Ap.J. Suppl., **199**, 25.

Rees, M. J. 1984, ARA&A, **22**, 471.

Rest, A., et al. 2005, Nature, **438**, 1132.

Rest, A., et al. 2008a, ApJ, **680**, 1137.

Rest, A., et al. 2008b, ApJL, **681**, L81.

Rest, A., et al. 2011, ApJ, **732**, 3.

Rest, A., et al. 2012, Nature, **482**, 375.

Rosati, P., Borgani, S., & Norman, C., 2002, ARA&A, **40**, 539.

Schaefer, B., & Pagnotta, A. 2012, Nature, **481**, 164.

Springel, V., Di Matteo, T., Hernquist, L. 2005, MNRAS, **361**, 776.

Targett, T. A., Dunlop, J. S., Cirasuolo, M. et al. 2013, MNRAS, **432**, 2012.

Tumlinson, J., et al. 2013, ApJ, **733**, 111.

van der Maaten, L.J.P. 2014, *Journal of Machine Learning Research*, 15(Oct):3221-3245.

Vogelsberger, M., Genel, S., Springel, V., et al. 2014, MNRAS, **444**, 1518.

Whitmore, B. C., et al. 2010, AJ, **140,** 75.

Williams, B. F., et al. 2014, ApJS, **215**, 9.

Williams, B. F., et al. 2015, ApJ, **806**, 48.

Williamson, M. S., Benson, B. A., High, F. W., et al., 2011, ApJ, **738**, 139.

Wright, D.E., et al 2015, MNRAS, **449**, 451.

Xu, B., Postman, M., Meneghetti, M., et al. 2015, ApJ, **817**, 85.

Zhu, G., & Menard, B. 2013, ApJ, **770**, 130.

9 Acronym and Jargon Dictionary

2MASS = 2 Micron All-Sky Survey (NASA funded ground-based survey)
AGN = Active Galactic Nuclei.
ALMA = Atacama Large Millimeter Array (ground-based radio observatory)
API = Application Programming Interface
ASASSN = All-Sky Automated Survey for Supernovae
CANDELS = Cosmic Assembly Near-Infrared Deep Extragalactic Survey (a multi-cycle *HST* treasury program)
CGM = Circumgalactic Medium
CLASH = Cluster Lensing And Supernovae survey with Hubble (a multi-cycle *HST* treasury program)
COS = Cosmic Origins Spectrograph (*HST* science instrument)
COSMOS = Cosmological Evolution Survey
CPU = Central Processing Unit
CSA = Canadian Space Agency
DES = Dark Energy Survey
DLA = Damped Lyman-α Absorber
DMS = Data Management System
DSMO = Data Science Mission Office
ELT = Extremely Large Telescope (any ground-based optical observatory with an effective aperture of 20 meters or more)
eRosita = All-sky X-ray space-based imaging survey in the medium energy range up to 10 keV (joint Russian and German mission)
ESA = European Space Agency
FDC = Flexible Data Center
FFI = Full Frame Image
FIRST = Faint Images of the Radio Sky at Twenty centimeters
FOS = Faint Object Spectrograph (*HST* science instrument)
Gaia = High-precision stellar astrometry mission (ESA)
GALEX = Galaxy Evolution Explorer (NASA mission)
Gb = Gigabit (10^9 bits)
GB = Gigabyte (10^9 bytes)
Gbps = Gigabits per second
GHRS = Goddard High Resolution Spectrograph (*HST* science instrument)
GHz = Gigahertz (10^9 cycles per second)
GPU = Graphics Processing Unit
HLA = Hubble Legacy Archive
HLS = High Latitude Survey (WFIRST)
HPC = High Performance Computing
HSC = Hubble Source Catalog
HST = Hubble Space Telescope (NASA/ESA mission)
HTC = High Throughput Computing
HTCondor = a workload management system for scheduling and prioritizing batch jobs on a compute cluster.
IDL = Interactive Data Language
IFU = Integral Field Unit
IGM = Intergalactic Medium
IR = Infrared (usually, wavelengths longer than 1 micron). Near-IR usually covers range 1 to 2.5 microns.
IRAS = Infrared Astronomy Satellite (NASA mission)
JWST = James Webb Space Telescope (NASA/ESA/CSA mission)
LLS = Lyman Limit System
LSST = Large Synoptic Survey Telescope (future ground-based facility)
MAST = Mikulski Archive for Space Telescopes
Mbps = Megabits per second
MOS = Multi-Object Spectrograph
NAS = Network Attached Storage
NASA = National Aeronautics and Space Administration
NIRISS = Near-Infrared Imager and Slitless Spectrograph (JWST science instrument)
NRAO = National Radio Astronomy Observatories
NVSS = NRAO VLA Sky Survey
OWL = Open Workflow Layer
PanSTARRS = Panoramic Survey Telescope And Rapid Response System
PB = Petabyte (10^{15} bytes)

PHAT = Panchromatic Hubble Andromeda Treasury (a multi-cycle *HST* treasury program)
ROSAT = Roentgen Satellite (X-ray observatory mission)
SAN = Storage Area Network
SDN = Software Defined Network
SDSS = Sloan Digital Sky Survey
SKA = Square Kilometer Array (future radio astronomy facility)
SN = Supernova (plural is supernovae, abbreviated SNe)
SOA = Service Oriented Architecture
SQL = Structured Query Language
SSD = Solid State Disk drive
STIS = Space Telescope Imaging Spectrograph (*HST* science instrument)
STScI = Space Telescope Science Institute
t-SNE = t-Distributed Stochastic Neighbor Embedding
TB = Terabyte (10^{12} bytes)
TESS = Transiting Exoplanet Survey Satellite (NASA mission)
UV = Ultraviolet (usually, the wavelength range between 100 nm to 350 nm)
VDI = Virtual Desktop Infrastructure
VLA = Very Large Array (ground-based radio facility)
WFC3 = Wide-Field Camera 3 (*HST* science instrument)
WFIRST = Wide-Field Infrared Survey Telescope (NASA mission)
WISE = Wide-field Infrared Survey Explorer (NASA mission)