Draft version May 12, 2019
Typeset using LaTeX **modern** style in AASTeX62

Petabytes to Science

Amanda E. Bauer,[1] Eric C. Bellm,[2,3] Adam S. Bolton,[4] Surajit Chaudhuri,[5] A.J. Connolly,[3] Kelle L. Cruz,[6,7,8] Vandana Desai,[9] Alex Drlica-Wagner,[10,11] Frossie Economou,[1] Niall Gaffney,[12] J. Kavelaars,[13] J. Kinney,[14] Ting S. Li,[10,11] B. Lundgren,[15] R. Margutti,[16] G. Narayan,[17] B. Nord,[10,11,18] Dara J. Norman,[4] W. O'Mullane.,[1] S. Padhi,[19] J. E. G. Peek,[17,20] C. Schafer,[21] Megan E. Schwamb,[22] Arfon M. Smith,[17] Alexander S. Szalay,[23,20] Erik J. Tollerud,[17] and Anne-Marie Weijmans[24]

[1]*Large Synoptic Survey Telescope (LSST/AURA)*
[2]*LSST*
[3]*DIRAC Institute, Department of Astronomy, University of Washington*
[4]*NOAO*
[5]*Microsoft Research*
[6]*Hunter College, City University of New York*
[7]*American Museum of Natural History*
[8]*Center for Computational Astrophysics, Flatiron Institute*
[9]*Caltech/IPAC*
[10]*Fermi National Accelerator Laboratory*
[11]*Kavli Institute of Cosmological Physics, University of Chicago*
[12]*Texas Advanced Computing Center*
[13]*National Research Council of Canada*
[14]*Google Inc.*
[15]*University of North Carolina Asheville*
[16]*Northwestern University*
[17]*Space Telescope Science Institute*
[18]*Department of Astronomy and Astrophysics, University of Chicago*
[19]*Amazon Web Services*
[20]*Department of Physics & Astronomy, The Johns Hopkins University*
[21]*Carnegie Mellon University*
[22]*Gemini Observatory*
[23]*Department of Computer Science, The Johns Hopkins University*
[24]*School of Physics and Astronomy, University of St Andrews*

Abstract

A Kavli foundation sponsored workshop on the theme *Petabytes to Science* was held 12^{th} to 14^{th} of February 2019 in Las Vegas. The aim of the this workshop was to discuss important trends and technologies which may support astronomy. We also tackled how to better shape the workforce for the new trends and how we should approach education and public outreach. This document was coauthored during the workshop and edited in the weeks after. It comprises the discussions and highlights many recommendations which came out of the workshop.

We shall distill parts of this document and formulate potential white papers for the decadal survey.

Keywords: Astronomy, Astrophysics, Work Force, Diversity, Inclusion, Software, Algorithms, Data Management, Computing, High Performance Computing (HPC), High Throughput Computing (HTC), Networking, Machine Learning, Cloud, Education, Management, Outreach, Workforce

Contents

1. INTRODUCTION

Contributors: *William O'Mullane <womullan@lsst.org>, Ting Li <tingli@fnal.gov>*

In the Petabyte era the lines between software, technology and science are blurred - the chance to do science with petabytes without major infrastructure is pretty slim. Therefore the importance of technology in science exploitation becomes ever more important, which also implies we pick up the pace in training the workforce and in the areas of education and public outreach.

The Kavli foundation sponsored a series of workshops on the theme *Petabytes to Science*[1], the second of which was held 12^{th} to 14^{th} of February 2019 in Las Vegas. The aim of the this second workshop was to formulate potential APC white papers. To facilitate this we discussed important trends, technologies, approaches to workforce management, education and public outreach. We took a holistic approach and built a single document encompassing several broad categories, namely:

- Science drivers (Section 3) - which science cases need new techniques and approaches?
- Data Management (Section 4) - what data management challenges does this present?
- Software (Section 6) - how should software be developed to meet those challenges?
- Technology and Infrastructure (Section 5) - what technologies and infrastructure is needed to under pin the services?
- Workforce and Inclusion (Section 8) - what training should we do to prepare ? How can we improve and diversify the workforce?
- Education and Public Outreach (Section 9) - through EPO can we increase awareness of the public about astronomy and ensure future finding streams? What are the challenges and opportunities for EPO?

From each of the sections a number of recommendations were identified, these are summarized in Section 2. For each recommendation we suggest the audiences they are useful to and the time period in which they should be executed (this may be seen as a sort of priority). The time periods or terms are short term (1-3 years), medium term (3-5 years) and long term (5-10 years).

The intention is to extract some decadal survey APC papers from these ideas. If you are interested in contributing to or endorsing white papers on these topics sign up here[2] or contact the authors listed in this document – names followed by an email address are the leads of each chapter.

Note: This document is a collection of ideas and a record from a workshop - it is not a polished document. We have made some effort to not have repetitions between chapters however we do not guarantee a pleasant coherent read.

[1] https://petabytestoscience.github.io/
[2] https://tinyurl.com/y2ksemp2

2. RECOMMENDATIONS

These tables summarize the recommendations in the document per audience we feel would be interested. Clicking the label or text will take you to the full recommendation in the document.[3] Please note that a recommendation may be aimed at multiple audiences and therefore may appear more than once in the tables below.

Table 1. Astronomer recommendations.

Recommendation	Area	Term
REC-1 Adopt common data models throughout the astronomical community	Data Management	Short
REC-3 Proprietary data time scales should be limited, and all datasets should be eventually made publicly available	Data Management	Short
REC-8 Improve long-term software and service support	Technology	Short
REC-11 Funding for sustaining core astronomical "community infrastructure" projects	Software	Medium
REC-12 Cultivating a sustainable research software ecosystem	Software	Short
REC-13 Create funding models and programs to support the development of advanced algorithms and statistical methods specifically targeted to the astronomy domain	Analysis	Medium
REC-14 Build automated discovery engines	Analysis	Long
REC-15 Promote interdisciplinary collaboration between institutions, fields, and industry	Analysis	Long
REC-16 Develop an open educational curriculum and principles for workforce training in both algorithms and statistics	Analysis	Medium
REC-17 Encourage, support, and require open publication and distribution of algorithms	Analysis	Short
REC-22 Software training as part of science curriculum	Workforce	Medium

Table 2. Manager recommendations.

Recommendation	Area	Term
REC-4 Long-term data preservation of datasets	Data Management	Long
REC-12 Cultivating a sustainable research software ecosystem	Software	Short
REC-15 Promote interdisciplinary collaboration between institutions, fields, and industry	Analysis	Long
REC-25 Recognize software as part of the career path	Workforce	Short

Table 3. University recommendations.

Recommendation	Area	Term
REC-22 Software training as part of science curriculum	Workforce	Medium
REC-26 Partnerships to support data science staff	Workforce	Medium

[3] In a Probability Density Function (PDF) it may be useful to note that CMD ← (CTRL on Windows/Linux) returns you to from whence you clicked.

Table 4. Agency recommendations.

Recommendation	Area	Term
REC-2 Eliminate barriers to public data access	Data Management	Medium
REC-3 Proprietary data time scales should be limited, and all datasets should be eventually made publicly available	Data Management	Short
REC-4 Long-term data preservation of datasets	Data Management	Long
REC-5 Develop a community wide architecture supporting Science as a Service	Technology	Long
REC-7 Enable support for full mission life cycle including long-term data products	Technology	Short
REC-8 Improve long-term software and service support	Technology	Short
REC-9 Fund cross-mission deployment	Technology	Medium
REC-10 Funding for software development in existing grant programs	Software	Long
REC-11 Funding for sustaining core astronomical "community infrastructure" projects	Software	Medium
REC-12 Cultivating a sustainable research software ecosystem	Software	Short
REC-13 Create funding models and programs to support the development of advanced algorithms and statistical methods specifically targeted to the astronomy domain	Analysis	Medium
REC-15 Promote interdisciplinary collaboration between institutions, fields, and industry	Analysis	Long
REC-16 Develop an open educational curriculum and principles for workforce training in both algorithms and statistics	Analysis	Medium
REC-17 Encourage, support, and require open publication and distribution of algorithms	Analysis	Short
REC-18 Programs to cultivate the next generation	Workforce	Long
REC-20 Long-term curation of materials	Workforce	Long
REC-21 Funding for innovative partnerships	Workforce	Medium
REC-23 Training activities and materials	Workforce	Short
REC-27 Support long-term technical capacity	Workforce	Medium

Table 5. Educator recommendations.

Recommendation	Area	Term
REC-22 Software training as part of science curriculum	Workforce	Medium

Table 6. Technologist recommendations.

Recommendation	Area	Term
REC-1 Adopt common data models throughout the astronomical community	Data Management	Short
REC-5 Develop a community wide architecture supporting Science as a Service	Technology	Long
REC-6 Enable new scales of research through data co-location	Technology	Medium
REC-8 Improve long-term software and service support	Technology	Short

REC-11 Funding for sustaining core astronomical "community infrastructure" projects	Software	Medium
REC-14 Build automated discovery engines	Analysis	Long

3. SCIENTIFIC CONTEXT AND DRIVERS

Contributors: *Adam Bolton <bolton@noao.edu>, Eric Bellm, Alex Drlica-Wagner, Ting Li, Raffaella Margutti, Gautham Narayan, Meg Schwamb*

Note: If you have come directly to this chapter we suggest you please read at least the Introduction in Section 1 before delving further.

The last two decades have seen a significant increase in the prominence of data-intensive, survey-scale astronomy. Surveys such as Sloan Digital Sky Survey (SDSS), Dark Energy Survey (DES), Pan-STARRS, and Zwicky Transient Facility (ZTF) have pioneered these modes. Even more ambitious projects such as Dark Energy Spectroscopic Instrument (DESI), Large Synoptic Survey Telescope (LSST), WFIRST, and Square Kilometer Array (SKA) are rapidly approaching, bringing new opportunities and challenges for petascale astronomical science.

From an experimental design perspective, the development of large survey projects and facilities has been driven by fundamental scientific questions about the Solar System, our Milky Way and its stellar populations and satellites, the evolution of galaxies and quasars, and the nature of dark matter and dark energy. These questions have in common the need to obtain significant statistics over large population samples or volumes, or alternatively, to realize significant probabilities for the discovery of rare objects or events.

Big surveys naturally lead to big datasets. These big datasets in turn bring qualitatively new challenges in data management, computing, software, and professional development that must be tackled to realize the scientific promise of the surveys themselves.

Big datasets from big surveys also open up diverse opportunities for *data-driven science*: research and discovery programs defined entirely on the basis of available datasets, not on the basis of collecting new data. This approach is especially empowering of exploratory science, as described in the series of Astro2020 white papers by Fabbiano et al. (2019a,b,c,d,e,f). Data-driven research can multiply the scientific impact of a survey, and can be especially effective for broadening participation in forefront astronomical research beyond those groups with the greatest access to resources. Data-driven science with large surveys calls on many of the same data-intensive methods as are required for "primary" survey science, while also presenting new challenges and requirements such as public data release and broad data accessibility.

In the following subsections, we outline some current scientific opportunities, and their associated data-intensive challenges, across a broad range of astrophysics and cosmology drawn from science white papers submitted to the Astro2020 Decadal Survey. In the subsequent chapters of this report, we address the crosscutting technology, methods, and professional considerations that will support success in these scientific areas in the next decade.

3.1. *Planetary Systems; Star and Planet Formation*

LSST will conduct a 10-year survey across the southern sky, revisiting the same locations approximately every three days. This time-resolved dataset will detect both transient objects in the fixed sky and moving objects in the Solar System. Chanover et al. (2019) describe the promise of LSST for the discovery of *dynamic* Solar System phenomena such as active asteroids and small-body collisions. To yield their scientific potential, these objects require rapid detection, alert, and follow-up observation. This implies the need for a coordinated real-time software infrastructure beyond the scope of LSST operations deliverables. Similarly, Holler et al. (2019) describe the Solar System science potential of WFIRST, which will require the deployment of robust moving-object detection algorithms within the WFIRST data management framework.

A core goal of the WFIRST mission is to conduct a microlensing census of extrasolar planets. Yee et al. (2019) and Gaudi et al. (2019) describe both core and ancillary science potential of this aspect of WFIRST, which highlights the algorithmic and software-systems engineering challenge of addressing diverse microlensing applications within a petascale dataset with quality comparable to space-based telescopes.

Ford et al. (2019) discuss the essential role of advanced statistical and machine-learning methodologies for optimal extraction of Doppler signatures of extrasolar planets with high-resolution spectroscopy in the coming decade. As the experimental forefront approaches the 10 $\mathrm{cm\,s^{-1}}$ precision necessary to detect true Earth analogs around Sun-like stars, new statistics and algorithms become ever more crucial.

3.2. *Stars and Stellar Evolution; Resolved Stellar Populations*

Pevtsov et al. (2019) highlight the potential scientific return for stellar astrophysics from digitizing historical astronomy data and making it available in accessible forms within modern data-management systems.

Dey et al. (2019) and Kollmeier et al. (2019) describe the potential for data-mining within large spectroscopic survey datasets (e.g. SDSS, DESI) to discover primordial Population III stars as well as new, rare, and unexpected stellar types.

Several Astro2020 white papers highlight the scientific potential that arises from combining multiple large-scale resolved stellar datasets. Asteroseismology results can be sharpened through the combination of time-series data from TESS, PLATO, and WFIRST with stellar spectroscopic parameters measured by SDSS-V, Maunakea Spectroscopic Explorer (MSE), and other surveys (Huber et al. 2019). Our knowledge of the structure, formation, stellar populations, and cosmological context of the Milky Way will be maximized through the combination of photometry, astrometry, and spectroscopy from multiple survey missions (Sanderson et al. 2019; Williams et al. 2019). Joint time-resolved analysis of photometric, astrometric, and spectroscopic survey data will also enable diverse astrophysical applications of stellar multiplicity (Rix et al. 2019). For all these scientific goals to be realized, full interoperability and combined analysis at the scale of millions to billions of stars will be required across all relevant surveys, which poses significant challenges in data management and software systems, as described by Olsen et al. (2019).

3.3. *Compact Objects; Time-Domain and Multi-Messenger Astrophysics*

Graham et al. (2019) highlight the explosive-transient discovery-space potential for LSST combined with Extremely Large Telescope (ELT) follow-up, which will only be realized if detection, filtering, and follow-up can be triggered rapidly by scientifically tuned software systems. Kirkpatrick et al. (2019) argue for increasing the transient and variable science return of the NEOCam mission through investment in the software infrastructure needed to detect, monitor, and alert on non-moving (i.e. non-Solar system) variable sources. This is an example of how software alone can qualitatively change the scientific opportunity space of a given survey/mission.

Cowperthwaite et al. (2019) argue for the importance of "target-of-opportunity" observing with LSST to follow up on LIGO gravitational wave triggers in search of a counterpart. This points to the need for sophisticated real-time data-management software systems such as LSST's to be implemented in ways that are flexible to the development of new and potentially unanticipated operational modes.

Binary systems with compact-object components provide "astrophysical laboratories" that will be central to many time-domain and multi-messenger applications in the coming decade. Maccarone et al. (2019) describe the scientific potential for increasing the sample of known stellar binaries with a black hole component, and highlight the importance of time-domain photometric surveys such as LSST, ZTF, ATLAS, PanSTARRS for identifying candidate systems through analysis of light curves to identify ellipsoidally modulated binaries and optically-outbursting X-ray binaries. Eracleous et al. (2019) describe the role of LSST and ZTF in catching the disruption of white dwarf stars by a black hole companion, which can be further informed by LISA observations when available. Littenberg et al. (2019) and Kupfer et al. (2019) describe the importance of LSST, ZTF, Gaia, BlackGEM, SDSS-V, and DESI for identifying ultracompact binaries that will be potential future persistent gravitational wave sources for LISA. In all these cases, algorithmic time-domain analysis and discovery implemented through science-driven software systems will be essential.

Palmese et al. (2019) highlight the potential for large spectroscopic surveys to provide redshifts for hosts of future gravitational-wave inspiral sources, both "bright" and "dark". These redshifts will enable "standard siren" cosmology in combination with the inferred intrinsic parameters of the Gravity Wave (GW) sources. This points to the need for spectroscopic surveys to make their data archives fully accessible.

Cutting across several of the scientific topics above, Chang et al. (2019) provide an overview of "cyberinfrastructure" needs for multi-messenger astrophysics in the coming decade.

3.4. *Galaxy Evolution*

Large extragalactic surveys and their associated data archives are a key resource for advancing our understanding of galaxy evolution. Behroozi et al. (2019) highlight the importance of large surveys, accessible data archives, and open software for advancing our knowledge of galaxy evolution through the particular method of empirical modeling.

Dickinson et al. (2019) envision a future spectroscopic galaxy survey that would provide a highly complete SDSS-like sample across multiple redshifts, which would enable a comprehensive study of the coevolution of galaxies and their stellar populations with the formation of dark matter halos across cosmic time. Additional galaxy-evolution (and cosmology) science drivers are discussed below in Section Section 3.6 in the context of combining data from multiple surveys and facilities.

Multiple quasar science opportunities in the next decade will be driven by survey-scale and data-intensive methodologies. Shen et al. (2019) describe the role of large, time-resolved spectroscopic surveys to map the structure and growth of quasars through the method of reverberation mapping. Fan et al. (2019) highlight the prospects for data mining in LSST and WFIRST for large samples of high-redshift luminous quasars, which can probe the coevolution of galaxies and their central SMBHs at early times. Pooley et al. (2019) highlights the role of LSST as a resource for discovery of strongly lensed quasars which can uniquely probe the dark-matter fraction in the lensing galaxy, while Moustakas et al. (2019) describe the role that these same systems can play in reconstructing the detailed structure of quasars themselves.

Lehner et al. (2019) highlight the importance of high-quality spectroscopic reduction pipelines and accessible data archives to maximize the science potential of high-resolution spectroscopy on large ground-based telescopes to trace the evolution of the intergalactic and circumgalactic medium over cosmic time.

3.5. *Cosmology and Fundamental Physics*

With its goal of understanding the contents and evolution of the universe as a whole, cosmology has driven many of the recent advances in "big data" astronomy. This trend is likely to continue through the 2020s. Many Astro2020 science white papers describe planned and proposed missions for which robust data-processing and data-management systems will be essential baseline requirements. These projects require not just basic management of petascale data, but also the automated execution of sophisticated inference algorithms—for galaxy shapes, photometric and spectroscopic redshifts, selection functions—across their entire datasets.

Slosar et al. (2019a) provide a broad overview of prospects for ongoing study of dark energy and cosmology with large-scale surveys. Dore et al. (2019) give an overview of the cosmological science capabilities of WFIRST via the channels of weak lensing, galaxy clustering, supernovae, and redshift-space distortions. Wang et al. (2019) describe the dark-energy science potential of multi-tracer wide-field spectroscopic surveys that achieve higher completeness and spatial density than existing or planned surveys. Slosar et al. (2019b), Ferraro et al. (2019), and Meeburg et al. (2019) describe the prospects for constraining models of inflation and early-Universe physics through the signatures of primordial non-Gaussianity in large-scale structure surveys. Pisani et al. (2019) describe the potential to constrain dark energy, neutrinos, and modified gravity in cosmic voids within densely sampled redshift surveys. Dvorkin et al. (2019) describe the prospect for measuring the absolute neutrino mass scale through several large-scale observational channels. Rhodes et al. (2019b) envision the definitive large-scale structure survey to map the three-dimensional position of all galaxies and dark-matter halos in the visible universe. Geach et al. (2019) envision a future wide-field spectroscopic survey in the sub-millimeter which would cover redshifts 1–10.

Large surveys and their associated software and data systems will likewise be central to the quest to understand dark matter in the coming decade. Gluscevic et al. (2019) describe the potential for galaxy and Lyman-alpha forest surveys, in combination with modeling and simulation of baryonic effects, to constrain the nature of particle dark matter. Bechtol et al. (2019) describe the dark-matter science potential of LSST on its own and in combination with spectroscopic facilities. Other channels for constraining particle dark matter with large spectroscopic surveys of galaxies and Milky Way stars are described by Li et al. (2019). Grin et al. (2019) describe how a combination of Cosmic Microwave Background (CMB), optical, infrared, and gravitational wave observations will contribute to our understanding of ultra-light dark matter candidates.

Survey-scale cosmological science in the 2020s will also leverage machine learning (ML) supported by large and well-calibrated datasets. Ntampaka et al. (2019) describe recent applications of ML to a diverse range of applications in cosmology, and highlights some of the most significant opportunities for ML to increase the scientific return from LSST, SKA, and other major future projects.

3.6. *Combining Multiple Probes of Cosmology and Galaxy Evolution*

A central theme of many Astro2020 science white papers at the interface of galaxy evolution and cosmology—and one that will significantly drive requirements for the computing, data, and software systems of the 2020s—is the need to combine and co-analyze data from multiple major surveys. These use cases imply requirements for data accessibility, interoperability, and mobility between data-hosting locations. They will also drive the astronomy and cosmology communities to leverage the capabilities of research-supercomputing and commercial-cloud computing providers in new ways.

Newman et al. (2019) and Mandelbaum et al. (2019) describe the synergistic potential for deep and wide-field survey spectroscopy to enhance the dark-energy science return from LSST. Chary et al. (2019), Eifler et al. (2019), and Rhodes et al. (2019a) describe joint analysis approaches for LSST, Euclid, and WFIRST that would enhance the resulting weak-lensing and galaxy-clustering cosmology measurements of these missions. Capak et al. (2019) describe the scientific benefit from coordination of "deep field" regions across multiple surveys and multiple wavelengths. (Here, standardized data and metadata formats will be necessary not only to realize the scientific potential of diverse datasets in common areas of sky, but also to enable discovery of existing datasets and coordination of planned future deep-field campaigns.) Furlanetto et al. (2019), Cooray et al. (2019), and Cuby et al. (2019) describe the potential for combining galaxy surveys (space and ground-based), 21cm surveys, and other probes to obtain a more detailed picture of the epoch of reionization.

Mantz et al. (2019) describe the importance of combining multiple large surveys across wavelength for the selection of uniform and significant samples of high-redshift galaxy clusters.

4. DATA MANAGEMENT

Contributors: *Anne-Marie Weijmans <amw23@st-andrews.ac.uk>, JJ Kavelaars <JJ.Kavelaars@nrc-cnrc.gc.ca>, Surajit Chaudhuri, Vandana Desai, Jamie Kinney, William O'Mullane, Alex Szalay*

Note: If you have come directly to this chapter we suggest you please read at least the Introduction in Section 1 before delving further.

In this section we recognize two of the main challenges related to data management in the next decade:

- Big Data: datasets will be of such large volume, that moving them across individual data repositories is not practical. This will affect the way that we interact with data, and has the risk that some users will be excluded from access to large datasets. (See also Section 5.2)
- Time Domain: datasets will contain a time domain element, i.e. will contain data of the same part of the sky obtained at different time intervals. This will put challenges on current visualisation and discovery tools.

To address these two challenges, we make the following recommendations:

REC-1 Adopt common data models throughout the astronomical community.
Area: Data Management. **Audience:** Astronomer, Technologist. **Term:** Short
The astronomical community should work towards a common data model. This will allow astronomers to concentrate on scientific exploration of datasets, without having to worry about data formats and structures

REC-2 Eliminate barriers to public data access.
Area: Data Management. **Audience:** Agency. **Term:** Medium
Astronomical public datasets should be accessible to everyone, and everyone should have the opportunity to contribute to astronomical public datasets

REC-3 Proprietary data time scales should be limited, and all datasets should be eventually made publicly available.
Area: Data Management. **Audience:** Agency, Astronomer. **Term:** Short
To maximize scientific output, and allow wider-community access of centralized funded projects, all astronomical datasets should be made publicly available and accessible after an appropriate but short proprietary time limit

REC-4 Long-term data preservation of datasets.
Area: Data Management. **Audience:** Agency, Manager. **Term:** Long
Long-term data preservation and management should be an integral part of community-wide project planning

We discuss these recommendations in more detail in the sections below.

4.1. *Interoperability*

In the 2020s, new datasets such as LSST, WFIRST, and Euclid hold enormous science promise. Realizing this potential for transformative science presents a number of challenges for data management. As we outlined above: the main challenges are the volumes of the data, as well as the additional dimension that time domain observations will bring to these large datasets.

4.1.1. *Common observation models*

Data centers should adopt a common observation model (REC-1). A common observation model is a set of standard metadata parameters that can be used to describe any astronomical dataset. The widespread adoption of a common observation model has many advantages, outlined below.

The large volume of data in the 2020s implies that re-processing these data will incur high cost, thus increasing sharply the importance of a common data model that can serve as the basis for information exchange and reuse. International Virtual Observatory Alliance IVOA is on the way to adopting the Common Archive Observation Model (CAOM) for images. It has already been adopted by a number of large archives, including the Canadian Astronomy Data Centre (CADC), the European Space Astronomy Centre (ESAC), the Mikulski Archive for Space Telescopes (MAST), and the NASA/IPAC Infrared Science Archive (IRSA).

Effective re-use of data requires careful, ongoing curation of this metadata model. This includes both preserving the expertise and context of what the nuances of a particular dataset are, but also periodically updating metadata to conform to new standards and meet new use cases. For example, astrometry of old datasets may need to be updated to support real-time querying/matching/aligning/jointly processing for time domain studies.

4.1.2. *Data storage*

Data Centers should adopt industry standards for data storage when possible, see also Section 5.2. Perhaps the most obvious challenge is simply storing the data. The large volumes mean efficiency in storage representation is important.

We recommend that data centers leverage ‘off-the-shelf’, open data management services, tools, and technologies that have been developed by industry. Moving to industry standards for things like images allow us to leverage new technologies such as the ability to stream and operate remotely on *objects* using standard tools. File systems as we know them will not be the most appropriate storage model at petascale levels. Alternatives include the use of cloud object stores, cloud compute, ‘big data native’ formats such as Apache parquet and OpenEXR, and cloud-optimized FITS (see cloud optimized GeoTIFF as an example https://www.cogeo.org). Traditional astronomy file formats (e.g. FITS) should be used as they were originally intended, for transport only. That being said, one big advantage of FITS files is their ability to co-package meta-data, while e.g. for Parquet there are only limited options to have meta-data included directly with the data. Data and meta-data should be managed together to not lose efficiency in analysis performance.

4.1.3. *Eliminating file systems*

The community should develop a flexible suite of application program interfaces to abstract the file system.

The previous recommendation calls for using storage formats that are optimized for the cloud, in order to meet the challenge of "Big Data" storage. This also implies that the current often used practice of storing files and file systems locally on astronomers' laptops of analysis will have to change to this more global approach of accessing and analyzing data remotely. To avoid a difficult transition for many individual astronomers, global file structures and formats should be abstracted by two layers of application program interfaces (APIs). The bottom layer consists of a limited set of (Virtual Observatory (VO))-based APIs implemented by data centers. We recommend that data centers implement a critical set of core VO APIs, including cone search, image search, spectral search, Table Access Protocol (TAP), the standardized language used to report observations of astronomical events VOEvent, and Ephemeris Lookup (still to be adopted by the IVOA). Other VO standard protocols have become obsolete, and should not be implemented (e.g. VOSpace in favor of S3). The top layer consists of user-facing APIs developed by the community to "hide" the file formats from the user. In the 2020s, this top layer should focus on Python. However, lightweight APIs can be built in other languages as community needs dictate.

4.1.4. *Interoperable science platforms*

Data Centers should provide a set of interoperable science platforms.

A science platform provides users with access to compute and analytic services close to the datasets of interest. With new astronomy survey datasets measured in petabytes, it is quickly becoming infeasible to copy entire datasets to another location for analysis. At the same time, it is increasingly common for researchers to leverage big data and inefficient parallel compute technologies to analyze large subsets, if not entire datasets. Cloud services provided by commercial organizations and Department Of Energy (DOE)/National Science Foundation (NSF)-funded High Performance Computing (HPC) centers offer both the scale of compute resources and the networking infrastructure required to analyze these datasets using modern techniques. Furthermore, by physically co-locating datasets, we make it possible for researchers to conduct investigations that incorporate data from multiple mission archives. Therefore, we recommend that the DOE, NSF, and other funding agencies encourage data archives to be physically stored and perhaps co-located in facilities which are accessible to the global research community and which provide the compute and higher-level analytical services that will be used to analyze these datasets at scale.

4.2. *Lowering the barriers to public data access*

Projects should eliminate barriers to public data access (REC-2), and limit the proprietary data time scale (REC-3). To make maximal use of astronomical datasets, every astronomer should have access to these datasets, and have the tools available to exploit their scientific richness. In the sections below we make suggestions that projects should adopt to ensure that barriers to work with data are removed: we concentrate here on astronomical community (including students), and refer to Section 9 for promoting astronomical data with the general public. We also recommend that although proprietary data has its use within the astronomical community (e.g. ensuring that the astronomers and students who invested in collecting and reducing the data have the opportunity to explore the datasets for their science), that these proprietary times are kept short to maximize over-all science output.

4.2.1. *Computational resources*

Projects should make their datasets available for remote data analysis As mentioned in the previous section, the big datasets of the next astronomical surveys will be too large to download and store on individual astronomers computing systems (laptops). Projects should therefore ensure that their data is available for remote analysis, and provide opportunities for cloud computing. This will ensure that the whole astronomical community will have access to the data, and that lack of large data storage and/or computing facilities will not prevent astronomers from taking part in the scientific exploration of large datasets. We note that cloud computing does require reliable internet connections, which for most of the astronomical community will be available, but not necessarily for a more general audience (e.g. schools and individuals in remote areas).

4.2.2. *Documentation*

Projects should allocate sufficient resources and attention to capturing the expertise on collection, processing and interpretation of their data products. A dataset is only as strong as its documentation. Without documenting the expertise needed to work with a dataset, scientific analysis based on that data has a high risk of being flawed. Including detailed documentation with data resources is therefore a must. The documentation that captures the projects expertise should be easily accessible: the documentation should be released at the same time as the datasets. The documentation should be clearly written, with jargon explained and with tutorials and examples for clarification. The documentation should not be aimed at the experts within a project, but be written with inexperienced, new users in mind (e.g. students). There should be a mechanism (e.g. helpdesk, forum), in place to collect feedback and errata, and the documentation should be updated and improved accordingly during the life time of the project. Having excellent documentation does not only lower the barriers of entry to work with large datasets, but will also be invaluable when the project has reached the end of its lifetime, and the datasets will (eventually) go into long-term archiving (see Section Section 4.3).

4.2.3. *Professional training*

Training resources on the exploration of large public datasets should be made available for free and on-line. To lower barriers for entry further, there should be training resources available for the astronomical community, to ensure that they can explore the richness of large public datasets. These training resources, such as tutorials, demos and notebooks, should be aimed at appropriate levels, as the education needs of a beginning students are different than those of a postdoc or faculty member. These resources should be available and accessible to a large audience, and therefore should be linked to from dataset documentation pages.

4.2.4. *Education and Public Outreach*

Data facilities should invest in collaborations with Education and Public Outreach teams. Having real astronomical public data available for education and public outreach purposes is a big advantage for developing resources that closely mimic and can even contribute to scientific research. As outlined in Section 9 the Education and Public Outreach (EPO) chapter of this document, we recommend supporting dedicated education and outreach groups with relevant expertise to maximize the impact of EPO activities. To work closely with these EPO teams, we recommend that each data facility has at least one team member to liaise with the EPO team, and provide input on data requirements for EPO activities.

4.3. *Long-term data preservation*

Long-term data preservation and management should be an integral part of community-wide project planning (REC-4) Data that is actively used will continue to exist in the community: there is a sort of Darwinian selection going on constantly. Expertise is therefore also kept reasonably current while the data are in use. But the implication is that some data will be getting used less and less over time, and at some point is going to be compressed (including documentation and possible email archives) and put into cold storage for long term preservation. Catalogs and derived data products could potentially persist in regular use for longer than their source data.

The long term preservation of data is a problem that is not unique to Astronomy or Science, neither in volume nor in characteristics. Such preservation has the following components:

1. Ensuring data integrity (no tampering)
2. Sufficient redundancy so that there is no single point of failure, to ensure data access
3. Packaging of information that provides "recoverability" of essential information
4. Funding for such preservation, as well as data format and software maintenance.

The first challenge (data integrity) is a general problem, and there are many techniques that have been developed in research and in industry to ensure integrity. These include tamper-proof logs and signature based comparison of multiple copies of preserved data including watermarking. We should select a preferred method in astronomy.

The second challenge is met b and perhaps co-located having multiple sites to ensure that there is no single point of failure. There is a cost vs. "how many failures you can tolerate" trade-off. Offloading this task to multiple vendors of public cloud companies is probably the simplest solution. One compelling reason to do that is because they will, due to market pressure, continue to support changing data formats and media as technology change. Through all these, the data should remain accessible, even when in cold storage.

The third challenge, which is packaging of information, is most critical and this is where unique aspects of Astronomy are relevant. A data dump in itself is not easy to interpret, especially after several years, when the experts that generated and worked with the data have moved on to other projects. Therefore, it is critical to have good documentation and metadata enrichment. We need to capture the *expertise* so we may want to compress and store communications such as Slack channels, mailing lists and logs etc. As well as the actual data. By having such additional catalogs, the "recoverability" of value from preserved data is much enhanced. However, we need to acknowledge that such often more informal and unsorted communication does not replace the need for comprehensive and understandable documentation and tutorials to work with the data. The value of archived communication would be for the (hopefully) rare instances that an issue occurs that is not documented probably, but was discussed on communication channels, and for historic and/or social studies.

Last but not the least, there is the funding question. There are two possible models. First model is to attach a "service fee" to every funded project to support ongoing high quality documentation. Alternatively, funding may be requested as we near end of the project. The payment model, especially to cloud providers, could be fashioned like what is done for title insurance for home purchases – an one-time payment for a fixed number of years

5. TECHNOLOGY & INFRASTRUCTURE

Contributors: *William O'Mullane <womullan@lsst.org>, Niall Gaffney <ngaffney@tacc.utexas.edu>, JJ Kavelaars, Frossie Economou, Surajit Chaudhuri*

Note: If you have come directly to this chapter we suggest you please read at least the Introduction in Section 1 before delving further.

We discussed many of the challenges and potential technological and infrastructure innovations which could be future solutions to current problems. In these discussions, we concluded that the goal was not to predict future problems and unknown technological solutions, but to unite and align cross mission and research community cyberinfrastructure needs. We should standardize and establish best practices based on those already found within missions. These goals are best served with a design that enables common user identity models, along with common data, software, and infrastructure as services joining systems in a loosely coupled cyberinfrastructure. This will drive the the field towards a more interoperable cross mission cyberinfrastructure by design rather than common API and piecemeal translation layers which we currently have. This will enable developers to reach velocity more rapidly as they move between projects and missions since they will be more familiar with the common development practices and reference architecture.

5.1. *Commodity services and software based community architecture*

The astronomy and astrophysics community community have historically relied on the development and use of bespoke software and hardware infrastructure to solve challenges related to the managing and analyzing datasets at a scale that was difficult to find in industry or other scientific domains. These requirements are no longer unique and we have access to a wealth of open source software, commodity hardware, and managed cloud services (offered by commercial providers and federally-funded institutions) that are well positioned to meet the needs of astronomers and astrophysicists Momcheva et al. (2019); Bektesevic et al. (2019). By providing documentation and reference implementations of the "astronomy stack" using these technologies and making it easier for researchers and missions to access cloud computing services, we can reduce operations costs, accelerate time to science, and increase the scientific return on Federally-funded research in astronomy and astrophysics.

Such an architecture/system will provide access to new technologies for improved data interoperability. For example to enable a system to recognize transients in multi-observatory data with more than just the photometry. By housing such data as observing conditions, instrument bias, and even observation proposals within the system, developers can implement common layers at higher levels to provide common access missions. This can be done without having to specify the complete system for gathering, managing, and formatting the data. Missions can enforce access to either sensitive or proprietary information through role based access control to the data. With a well designed service oriented Cyberinfrastructure,

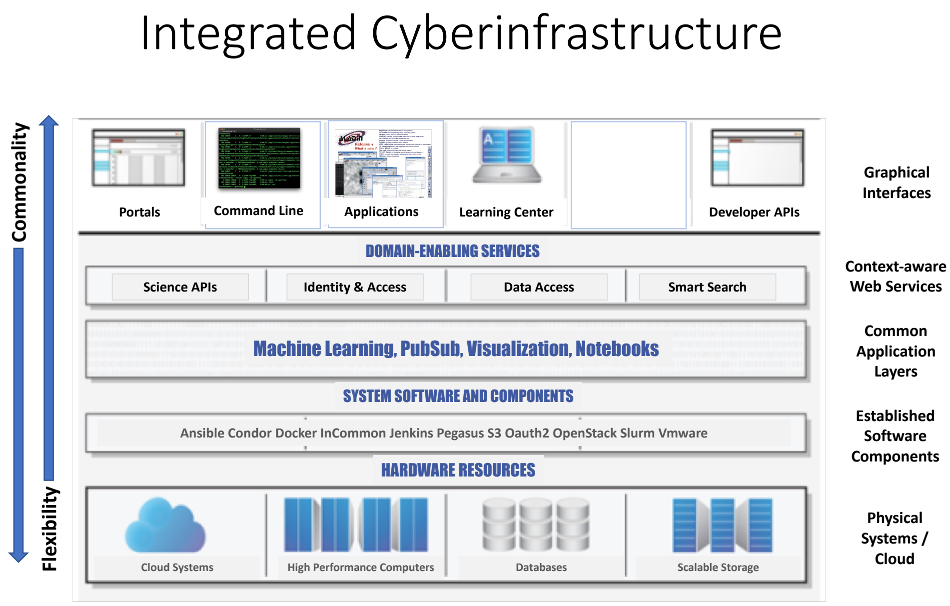

Figure 1. An example a cyberinfrastructure built on an Infrastructure as Code design model. Note that while this example does not have astronomy-specific tooling, our recommendations highlight the importance of developing astro-specific layers that are fully accessible to scientists in both the application and the graphical interface layers. Figure 2 presents an LSST/Astronomy instantiation of this.

cost can be minimized as less coordination will be needed to implement cross mission services.

REC-5 Develop a community wide architecture supporting Science as a Service.
Area: Technology. **Audience:** Agency, Technologist. **Term:** Long
Agencies should fund the major missions to define and adopt a community wide supported data and compute service architecture with easy to adopt components leveraging widely adopted infrastructure standards both in the community and in industry. This "Infrastructure as Code" (Morris 2016) approach lowers the bar to entry and allows for easier adoption of more standardized services that will enable large-scale astronomical research in ways that are well demonstrated in plant genomics (CyVerse and Galaxy), natural hazards (Designsafe), and surface water research (Hydroshare).

Many research communities have accelerated their time to discovery and lowered their cost of integration by adopting a common community wide architecture that is supported by multiple data and computational service providers. While attempts prior to the past decade have been moderately successful, the current shift in development across industry to the support of smaller services rather than monolithic data and compute systems allows for faster and more cost effective deployment across communities. By encouraging the definition and production of an astronomy focused, community wide reference architecture, perhaps by changing the funding structure, we can being to have a menu of services easily implementable across service providers. Design and support for this infrastructure should be community driven, prioritized and funded, to allow for development of features across missions and science use cases.

Pictured in Figure 1 is the structure of a cyberinfrastructure (CI) that has been used across multiple fields from plant and animal genomics (CyVerse) to natural hazards engineering (DesignSafe). This shows the layers of the CI from the interfaces for service access exposed at multiple levels, the common domain wide enabled services, and a collection of system level components that support the higher levels of the CI. The lower down the diagram are commodity layers based on well established and supported components. As one moves up from these layers, more abstraction can be done to expose these pieces in domain or even question level interfaces. By making these abstractions, more universal service can be developed that can be applied more globally across the entirety of the cyberinfrastructure as a whole. An example of this would be authentication, where each university or agency may provide their own authentication method but unifying services like CILogin can bring those together to give global spaced identity for a wide range of users based on disparate authentication systems. By providing this structure along with a reference architecture of these System Services based on well supported software components, providers are easily able to both deploy and support these common services which enable cross mission and center interoperability. This structure also reflects how this architecture allows for greater reusability as one gets closer to the actual implementation of these services while supporting greater flexibility and general usability as one works further from the core components. This service architecture should be based on using standard reusable software from many of the established standards developed outside of astronomy (e.g. common authentication mechanisms such as CILogin, standard data and metadata management systems). Standard API interfaces should also be used to expose these components to higher level APIs. Data formatting and metadata structure can be exposed at the service level, allowing for more data and metadata reuse.

Such an architecture should be developed in a cloud and vendor agnostic manner. When needed, vendor specific software or cloud service can be integrated by a mission, but by isolating them in the cyberinfrastructure at the lowest level, their potential impact on the overall system is minimized. When possible, standard interfaces should be used to abstract out these differences (e.g. standard object store access like S3, standard database interfaces like ODBC) and should be reflected in the reference architecture documentation. Where practical, computational environment abstraction layers such as container technologies (e.g. Docker) should be used to associate each applications computational requirements to the application rather than having to enforce upgrades and updates across the complete infrastructure. Where specific hardware environments are required (e.g. Google's Big Table or Tensor Processors), it must be required that the interface to these services leverage common API or software layers for access (e.g. SQL or Tensorflow) to allow for simpler migration to future or separate vendor's systems.

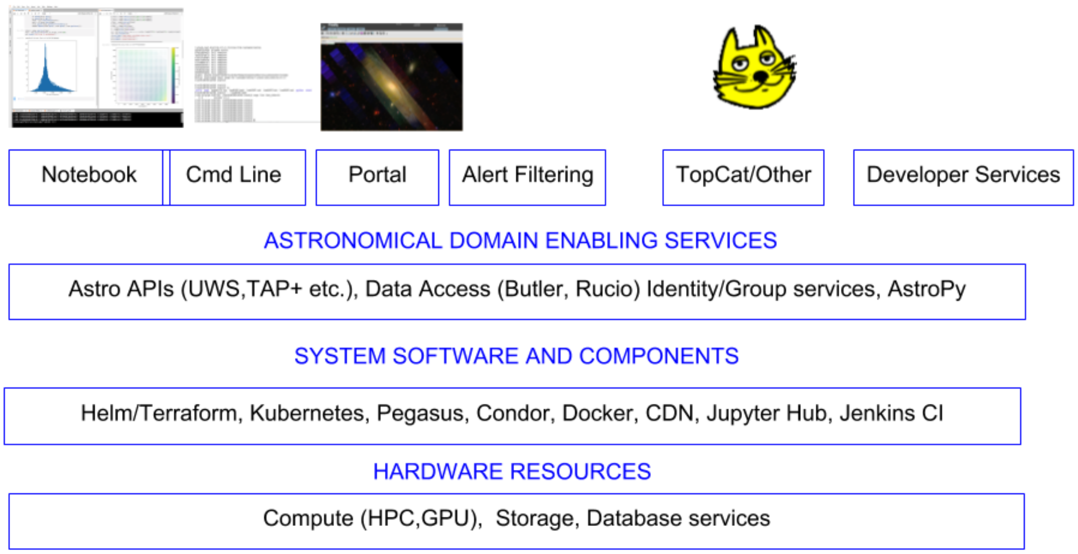

Figure 2. An example LSST cyberinfrastructure built analogous to the CI model shown in Figure 1.

Many of the advances in the current data revolution have come about from the broad adoption of commodity hardware and software services being applied in domain agnostic ways. While over a decade ago, astronomy was one of the first domains to explore these technologies, the current momentum in the area is driven outside of the astronomical field. Current technologies like Spark, Cassandra, Kubernetes, and MongoDB were all created to support large data problems outside of astronomy but are finding support in the astronomical community in a piecemeal manner. By shifting the focus from local to a more distributed cyberinfrastructure, such new technologies could be implemented and leveraged much quicker than if each center had to support and integrate their own solution, Further, by embracing and adopting a more commodity base infrastructure will allow current and future projects to choose and experiment with hardware alternatives such as TPUs, FPGAs, or Quantum computing as they become more commonplace or to integrate newer architectures that best suited to the problem.

There is a current tension or incompatibility with the direction of High Performance Computing HPC to many more cores with not much more memory and large image processing which requires more throughput. Historically, systems designed for HPC were more suited to simulation as opposed to the embarrassingly parallel yet often memory intensive High Throughput Computing HTC. While HTC does not get as much attention, it has been key to missions such as the Human Genome project, LIGO, and SDSS. While some are moving to bridge this gap (some in support of missions such as LIGO and others because simulated data analysis is becoming as complex as observational data analysis), agencies should continue encourage and fund national computing centers to address the needs of both communities. Further, they should enable simulated datasets to coexist within the infrastructure with observational datasets to support the full cycle of astronomical discovery. Finally, they should formally adopt support for mission long computational support at these facilities for all major missions.

We note the general idea here especially on the importance of cloud are compatible with Rob Pike's thoughts[4].

Where ever possible, review of cyberinfrastructure driven proposals should be ranked on both their immediate impact to the field as well as their ability to sustain such impact through several technology and vendor cycles. Proposals should also be ranked based on their reuse and/or integration into the overall cyberinfrastructure developed across the astronomical community.

5.1.1. *Identity and Access Management (IAM)*

User identity is key to all data systems to date. In the past, each mission has had its own IAM system to provide data access for embargoed or otherwise restricted data access and for accessing services. With a more global CI, this problem must be further abstracted to allow for the notion of identity of a user (aka authentication) and permissions (aka authorization) for services and data. By separating these pieces, identity can be brought from multiple sources (CI-Login, OAuth, ORCID) while the permissions can be enforced by each provider. All aspects of this CI must embrace role based authorization for both data and services in a federated manor so that data and services can be effectively orchestrated while not impacting site or mission specific access restrictions. While the authentication may be global, each provider will enforce their own access roles for all users. As data analysis moves into the petabyte and even exabyte scale across multiple missions and multiple computational environments, it will be paramount to for all members of the cyberinfrastructure to share a common IAM infrastructure to allow for federated access controls for both data and computational services.

Proposals and missions should be ranked on their adoption of such an IAM for data and for software services. Further, by creating a common CI layer for identity, missions and smaller services will be able to adopt and adapt the common system to their needs, thus saving development costs for implementing and finally supporting their own system.

5.2. *Data co-location and creating data lakes*

REC-6 Enable new scales of research through data co-location.
Area: Technology. **Audience:** Technologist. **Term:** Medium
We must enable the co-location of data thus enabling large scale cross mission research. While some research are well supported using specific services, ones which require comingling data to produce new data products or results should be able to schedule access to co-located data in ways similar to acquiring new data.

Astronomy archives, especially in the USA, are somewhat fragmented. Though TB scale permanently co-locating all data n one or more centers is technically possible it is probably

[4] https://drive.google.com/open?id=1kYsavh900I2o6z1lPfjFFXamyBkoEdCW

not desirable nor socially possible. In the PB data era permanent co-location becomes less feasible, yet researchers will require data lakes or reservoirs to house massive datasets and allow computations to be done across multiple missions and epochs of data at scale. The data lake concept, where data from multiple missions and schemes are temporarily co-located so as to allow codes to more tightly integrate with multiple data sources, is the most attractive for researchers (who have become accustom to immediate data access), the idea of a reservoir that can be filled and drained of data based on the demand of users is one that will need to be explored due to the economical viability of any one institution housing all the data.

Collocation is more than just interoperability but will also mean generating new formats e.g. parquet files (e.g.) to make dynamic data frames and data services as demands change. It will also mean generating new data products across missions which are equally valuable for both preservation and publication. Thus the lake is more than a simple pool of storage, but should be operated similar to other key infrastructures in the observational astronomical infrastructure.

Software and technologies change on timescales faster than a decade (see also Section 6.6) and data centers need to be agile enough to keep up. One approach is to create interoperability interfaces to allow data to be pulled or pushed from managed repositories to dynamic data lake environments where users can produce their own custom subsets mixing the available datasets. Data movement will not be simple (see also Section 5.2.1) and to not be prohibitive specific infrastructure would need to be supported and evolved.

While the immediately obvious argument for co-locating data is the potential for scientifically rich co-processing of heterogeneous data holdings (e.g. Euclid, WFIRST and LSST), the advantages do not end there. Co-locating large data holdings on a common commodity computing platform enables bring-your-code-to-the-data capabilities (sometimes referred to as Science Platforms or server-side analytics) to enable co-analysis of data from a single service without necessitating data transfer. For example, a single Jupyter notebook can present an runnable analysis drawing on separate datasets (e.g. a multi-wavelength analysis of a class of objects). Furthermore, co-locating data holdings allows the co-location *and sharing* of services accessing those data holdings. That not only includes the possibility of collaboration in sharing a single API service between data publishers, but also reducing the development burden of infrastructural services (for example, documentation infrastructure could be shared with multiple missions, migration paths to new operating systems are easier, etc). In an era where Infrastructure as Code engineering paradigms represent emerging best practice, re-using the code that underwrites common astronomical services because they are developed within a common underlying infrastructure (such as Docker containers orchestrated by Kubernetes) provides an avenue for fruitful ongoing collaboration between data publishers and better value for money for science infrastructure dollars.

5.2.1. *Network layer*

One area where Infrastructure as Code does not work is networking. While data transfers can be better optimized depending on the nature of the data being transferred, network as a service will require optimizations often at the point to point level. Tools like PerfSonar and others from Internet2 can help optimize connections. But these have been used to optimize research done at the terabyte scale today. Where as in the TB scale it was possible sometimes to move the compute to the data (e.g. SDSS), discovery often comes from the TB scale data lakes. To meet the demand for petabyte scale data motion as a service in a terabit network age, systems to support Just In Time data delivery are needed for any large scale data collocation and must be a part of the support for the overall infrastructure of repository/cloud/data center collaborations. Computation on large datasets may need to be scheduled no different than any other instrument used in observational astronomy.

We recommend funding projects in astronomy (and in other research domains) to create the data management and migration layers and best practices that will enable these new forms of observation. We also recommend that proposals show how they will develop and sustain such services over the course of the mission and beyond.

5.2.2. *Vendor freedom*

When co-location of data holding and services on a cloud platform is discussed, inevitable there are concerns raised about "vendor lock-in". These objections are often rooted in a misunderstanding of the nature of these services. First of all, these services often themselves share common open source technologies between them (for example there are a number of object store technologies that implement the an interface compatible with Amazon's S3 service). But more than that, the commercial landscape is designed around low barriers to change: Google Cloud Platform tries to acquire Amazon Web Services customers by making its services easy to switch to. Moreover all these platforms drive a service-oriented architecture that inherently results in more portable systems. In any case if one is concerned about vendor lock-in, in-house data center infrastructures are the worst possible choice: the inevitably lead to infrastructure-specific choices that are poor candidates for evolution, and they typically lack the effort to support ongoing refreshing of the technical stack thus creating on-going support burdens and a change-averse culture. For example, LSST is a project that will go into operations in 2022, and yet LSST Data Management (DM) are frequently called on to support CentOS 6, an operating system released in 2011, because that is the only Operating System (OS) some university in-house clusters support for researchers.

5.3. *Operations*

REC-7 Enable support for full mission life cycle including long-term data products .
Area: Technology. **Audience:** Agency. **Term:** Short
Agencies should revisit the model separating funding and requirements for development and operations of large scale missions for which the data and services are key deliverables as well as steel and concrete. Such services are under continual development and integration and, in the current environment, can not simply be maintained in the same way physical facilities are.

A more operations oriented view of construction by funding organizations would lead to facilities which are more cost effective to run. Recognizing that software and cyberinfrastructure development work is distinct from concrete and physical facilities would also help to make more maintainable and agile cyberinfrastructure. Current MREFCs funding splits construction from operations thus limiting support on ongoing work for cyberinfrastructure, there is little incentive in construction to build easy to operate systems.[5] If we blur the line between construction and operations for cyberinfrastructure the issue becomes more one of long term support.

REC-8 Improve long-term software and service support.
Area: Technology. **Audience:** Technologist, Agency, Astronomer. **Term:** Short
Funding should support repositories not just for code and data, but for computational environments. Use of proven standards in the wider research community for sharing and discovering runtime-ready software using software container environments like Docker and DockerHub but with domain specific curation (e.g. BioContainers) is crutical[6] for both broader impacts of software products and result reproducibility as compute environments continue to rapidly evolve with low emphasis given to backward compatibility.

Long term support could be improved by requiring proposals to state how code reuse and common practices will be used. Looking for proposals which aim to push to the Code and API communities (AstroGit? AstroContainers? AstroHub) and which aim to build on common software development practices. Of course we should also foster development of those best practices based on best practices outside astronomy. Concretely proposals could have a line item for making software reusable in funding budgets and funding agencies should see that as a good thing and try to develop metrics for success in the area.

We must also consider how to move to a more self service architecture in astronomy such as GitOps (Limoncelli 2018) - that requires some rigor but establishment of adhered to best practices would be a start.

Needless to say all of the code should be available under open licensing such as Apache (Apache Public License (APL)) or Gnu (GNU Public License (GPL)) public license.

REC-9 Fund cross-mission deployment.
Area: Technology. **Audience:** Agency. **Term:** Medium
Missions that develop software that can and should be adaptable to other common goals in other missions should be funded to develop and support cross-mission resources as part of this and other cyberinfrastuctures.

Past examples of this success can be found as far back as the IRAF environment and include now significantly more broadly adopted numpy, the cost of development and support have been paid back multifold across many funded projects and missions which did not have to develop their own versions of software.

[5] This is also true for physical facilities e.g. autonomous operation is usually not a requirement.
[6] Critical and crucial https://www.urbandictionary.com/define.php?term=crutical

Deployability is part of the problem but service oriented architectures are and will remain at the forefront for at least the next decade. So we should now be thinking more of software as a service and defining infrastructure as a service. This would all funding agencies to push us more toward commodity compute and infrastructure services thus concentrating efforts on the astronomy problems at hand rather than the computer science problems.

Funding agencies could also favor proposals that use/leverage existing software solution, it may take time but this would be a positive fundamental change in astronomy software.

5.4. *Sustainability and effectiveness*

An architecture as laid out in Section 5.1 gives us a framework to start a sustainable software development in astronomy. Software sustainability is a complex issue, with mixed results in astronomy — see Section 6.6.2 for more on this complexity. The approach laid out here, however, presents greater challenges than ever before in sustainability, as larger datasets and more complex infrastructure means that reinventing the wheel will become ever more costly.

By providing a reference architecture we can allow for more openness, collaboration, and when necessary, competition in each component. While competition and alternatives can be useful in a controlled manner, i.e. funding two approaches for a specific component, by providing a reference architecture this can be adopted only when needed without interfering with other layers. We should also consider that this architecture may be good for a decade at which point it also should be revisited - such a refresh should be built in to our thinking from the start. This is critical for sustainability because sustainability is much more challenging if the architecture does not keep up with the technology, as more and more "hacks" become necessary until the house of cards comes toppling down.

The Astropy project has been successfully fostering a community approach to the user-facing end of Python astronomy. They do face challenges for funding and are beginning to tackle some management issue for an organically grown organisation. This has been successfully because they have worked hard to join the zeitgeist of open software and have dedicated individuals who believe this is a useful project - and many users who agree. The project also produces useful tools. The role of the Astropy project in the ecosystem can be misunderstood i.e. it is not meant to be a data management system or a replacement for a mission-specific toolchain, rather it is a toolbox that is accessible to astronomers for doing many of the tasks they want to use or understand. While the re-use and contribution to these tools by missions or observatories is desirable (see Section 6), without a clear understanding of where the components lie in a larger architecture, it is almost impossible for astronomers and projects to understand how they can fit such components into their system. A reference architecture can thus help define where projects like this belong and how they fit with other efforts. For example, LSST will not process data using `astropy` but will make their data products accessible and usable with particular Astropy project interfaces and tools. It has taken both projects a while to understand this delineation because neither LSST nor `AstroPy` had a clear reference model to work with.

6. SOFTWARE

Contributors: *Arfon Smith <arfon@stsci.edu>, Erik Tollerud <etollerud@stsci.edu>, Kelle Cruz, Stuart Mumford, Nirav Merchant, Gautham Narayan, Alex Drlica-Wagner*

Note: If you have come directly to this chapter we suggest you please read at least the Introduction in Section 1 before delving further.

In the Petabyte era, all projects are software projects — that is, sophisticated software is necessary throughout the system for essentially any scientific output at this scale. That said, the term software can mean many things to many people. Other Sections (e.g. Section 3, Section 4, Section 7) discuss the *content* of this software, and the software "infrastructure" components and how they fit together are discussed in more detail in Section 5. Here, by contrast, our focus is on the process by which software distributed to and used by the astronomy community is built and funded. We particularly focus on community software as it is the most relevant for the community to be involved in and for funding agencies to support. Note that that while clearly critical to the process of software development, relevant career and workforce issues are discussed separately in Section 8.3.

6.1. *Recommendations*

REC-10 Funding for software development in existing grant programs.
Area: Software. **Audience:** Agency. **Term:** Long
Software that enables science should be allowable as a sole deliverable for all existing funding programs (e.g. NSF AAG, NASA ROSES, postdoctoral prize fellowships). It should not be necessarily coupled to a specific science effort, as long as the software is of demonstrable use to the scientific community.

REC-11 Funding for sustaining core astronomical "community infrastructure" projects.
Area: Software. **Audience:** Agency, Astronomer, Technologist. **Term:** Medium
Funding agencies and the community as a whole should support funding of domain-specific community-developed software projects e.g. Astropy project, SunPy. Such projects should be recognized as vital infrastructure and placed on an equal footing to physical facilities such as national observatories. This support should be available for domain-specific software, rather than funding being primarily tied to interdisciplinary applicability. It should also be allowed to fund community-development efforts in addition to actual code development.

REC-12 Cultivating a sustainable research software ecosystem.
Area: Software. **Audience:** Agency, Manager, Astronomer. **Term:** Short
Funding agencies should include as part of their review criteria for all astronomy grant programs: 1) A plan for how software (not just data) will be managed to support the science of the grant, 2) How proposed software development fits into and supports the wider

ecosystem of available tools, and 3) Favor programs that propose developing community software as part of their funded activities. These same goals and considerations should also be considered and acted on by the broader astronomy science community when e.g. working as grant review panelists.

Note that cultivating a research software *workforce* is critical to all of the above. Hence, while not detailed in this Section, the recommendations of Section 8 are also as critical to the discussion in this section as the above.

6.2. *Why shared software matters, today and in the next decade.*

In this chapter, we argue that the Petabyte era of discovery in astronomy means that the role of software is increasingly important and that well-organized, well-maintained software serves to shallow the learning curve, enable scientific investigation, and lends confidence to scientific results. To set scope, though, we emphasize this mainly applies in the context of shared software. That is, the "throw away" analysis script written, say, by a graduate student when writing their thesis and never shared with anyone else does not count for this discussion. However, that changes when the same student shares that script with their collaborators, makes it available online, or contributes it to The Astropy Project, rOpenSci, or another open community software resource. Such an act makes the software part of a community process, and the astronomy community is the target of this discussion. Hence, in this chapter we are focused on *shared* software.

6.2.1. *Software is everywhere*

> "**Software is a central part of modern scientific discovery**. Software turns a theoretical model into quantitative predictions; software controls an experiment; and software extracts from raw data evidence supporting or rejecting a theory" - Gaël Varoquaux, scikit-learn[7] creator, (Varoquaux 2013; Pedregosa et al. 2011).

Software is an integral, and growing part of the scientific endeavor: It is responsible for driving the control systems of instruments, the operation of surveys, the processing of raw data products, the extraction of physical parameters, and the theoretical modeling of physical systems, software is critical to all parts of modern computational research. Indeed, 'Software is eating the world' (Andreessen 2011). This reality is well-recognized by the scientific community: In a survey carried out in 2009, more than 2000 scientists reported that software was either important or very important to their research, and that it would be impractical for them to carry out their research without it (Hannay et al. 2009).

The rapid increase in the size and complexity of astronomical experiments and the data they produce has led to an increasing demand for astronomical software. An illustration of this point is the LSST project and their allocation of 25% of the construction budget ($187M) for data management software, infrastructure, and services[8].

[7] https://scikit-learn.org
[8] https://www.nsf.gov/about/budget/fy2018/pdf/30b_fy2018.pdf

Over the last decade, in a large part driven by broader changes in the cultural 'norms' of modern software development and a shift towards open source software being 'the new normal' (LeClair 2016; Gnau 2017), the astronomical software environment has changed rapidly: Large experimental projects (such as LSST, JWST, DESI, DKIST) are writing extensive code bases and releasing these tools as open source software[9][10][11][12]. At the same time, individuals are becoming increasingly likely to distribute and share their code broadly with the astronomical community and mechanisms for publishing these software products have expanded as a result (AAS Publishing 2015; GitHub 2016; Smith et al. 2018; Astronomy & Computing 2013). In this data intensive future, where software permeates scientific investigation, it is critical that the contributions of software developers are recognized and that individuals are provided with the necessary resources to succeed.

6.2.2. *Software encodes knowledge*

As datasets become larger and our analysis methods more sophisticated, an increasing fraction of the scholarly method is expressed in software. This presents opportunities and challenges. One potential opportunity is that the 'centralization' of astronomy (i.e. the trend towards smaller numbers of large facilities, often with open datasets) means that any software built (and shared) leveraging these facilities has a higher reuse potential. A major potential risk, identified by others (Donoho et al. 2009; Yale Roundtable Declaration 2010), is that as the fraction of our research method is captured in software, if this software isn't shared (e.g. as open source), reviewed, or tested, the reproducibility of our science is increasingly at risk.

6.2.3. *Software for reproducibility*

As projects become increasingly complex, ever more discrete software components are combined to produce analyses and data products. However, despite this complexity, much of the code being used is not documented, let alone complete with unit tests that can validate performance. These shortcomings can have real-world consequences, as illustrated by the failed Mars Climate Orbiter mission, where software calculations were carried out assuming metric units, but navigation software was programmed assuming imperial units, leading to a premature and fiery end to the mission in the Martian atmosphere. While this is an extreme case, it is an illustrative bounding case for more subtle problems in analyses that lead to biases which are not detected. Such problems are surprisingly common even in the computational science literature (Collberg & Proebsting 2016), much less more applied fields like astronomy. While progress has been made in developing technologies to improve this (e.g. easy and widely available software repositories like GitHub, containerization technologies like Docker, etc), many of these technologies are still aimed at early-adopters in science rather than the mainstream.

6.3. *Progress in the last the decade*

[9] http://github.com/lsst
[10] http://github.com/spacetelescope
[11] http://github.com/desihub
[12] https://github.com/DKISTDC

Many of the issues highlighted in this chapter are not new. In particular, we highlight a white paper from the Astro2010 decadal review with a similar scope: Weiner et al. (2009). That paper discussed areas of concern and specific recommendations, some of which have improved materially, while others have seen little progress. We discuss the recommendations of that paper here to provide a historical context and guidance for the future.

1. Weiner et al. (2009): *"create a open central repository location at which authors can release software and documentation"*. Enormous progress in this area has been achieved in the last decade. open source software repositories, chief among them GitHub[13], have become a defacto standard for storing software in astronomy. The wider adoption of Python has improved the packaging and release process due to the Python Package Index[14] and the ecosystem of easy-to-host documentation tools that support Python, like Sphinx[15] and ReadTheDocs[16]. While these are not a perfect solution for some languages and science domains, the presence of a much larger and better-funded user base (open source industry software) has made them stable enough to be adopted for astronomy's use and can likely continue to do so for the foreseeable future.
2. Weiner et al. (2009): *"Software release should be an integral and funded part of astronomical projects"*. Progress in this area has been mixed. While large efforts for this decade like LSST, JWST, DESI or DKIST have large first-class software components, many smaller projects or individual grant-level efforts continue to treat maintainable or reproducible software as an afterthought to be dealt with in whatever time is left over by graduate students or postdocs rather than a necessary part of the scientific endeavour. While funding agencies like the NSF, DOE and NASA have required data management plans, there has been less progress on establishing firm requirements or expectations of sustainable software (although a recent NASA-driven consideration of these issues is available in National Academies of Sciences, Engineering, and Medicine 2018).
3. Weiner et al. (2009): *"Software release should become an integral part of the publication process." and "The barriers to publication of methods and descriptive papers should be lower."*. Considerable progress has been made in this area. The American Astronomical Society (AAS) journals now allow software-only publications on equal footing with more traditional science publications (AAS Publishing 2015), and other major astronomy journals like A&A and PASP do as well. New approaches to publication like the Journal of open source Software[17] (Smith et al. 2018) or the Astrophysics Source Code Library[18] are now providing alternate ways to publish software that are indexed in Astrophysics Data System (ADS). Software archives

[13] https://github.com
[14] https://pypi.python.org
[15] http://www.sphinx-doc.org
[16] https://readthedocs.org/
[17] http://joss.theoj.org
[18] http://ascl.net

like Zenodo[19] now connect with GitHub to make publication of software via DOI almost frictionless (GitHub 2016). While there are still challenges in identifying how software citation should work in these areas, tangible progress and recommendation is being made (Smith et al. 2016). The "cultural" elements of ensuring these publications are viewed with the same level of value as other publications may also be improving, although concrete data in this area is lacking. While somewhat less progress has been made in ensuring open software is a truly integral part of publication, the same resources noted above have made it much easier to preserve software long-term. More challenging is preserving the *environment* software has been run in. While technologies like Docker or virtualization provide a possible path, they have not been adopted widely across the community thus far, and represent a possible major area of development for the 2020s.

4. Weiner et al. (2009): *"Astronomical programming, statistics and data analysis should be an integral part of the curriculum"* and *"encourage interdisciplinary cooperation"*. While some progress has been made in this area, there are many challenges remaining. We defer further discussion of this to Sections Section 8, Section 9, and Section 7.
5. Weiner et al. (2009): *"more opportunities to fund grass-roots software projects of use to the wider community"*. While such projects have grown remarkably in the last decade (see Section 6.4.1, major challenges still remain in *funding* such projects in a sustainable manner, and these form the core of some of our recommendations.
6. Weiner et al. (2009): *"institutional support for science programs that attract and support talented scientists who generate software for public release."*. Some of the elements of this recommendation have grown with the advent of "Big Data" and "Data Science" academic positions in astronomy. There has also been a growing recognition of the importance of research-oriented software positions, particularly in Europe (e.g. Research Software Engineers International 2018). However, there are very few viable pathways for researchers who develop software of broad use as part of their research program if it is not considered a "hot" field. Because, as this book demonstrates, there are likely to be *more* areas where deep software expertise is critical to science in the coming decade, the need for the field to nurture such career paths will only become more acute. Hence this is also a key element of our recommendations.

There is one final distinction to be highlighted relative to the last decade: it is clear that software has become more mission-critical than in the past. As the other chapters of this book highlight, in the coming decade(s) large-scale science will require larger and more complex software. These generic concerns about software development are therefore multiplied across the deeper layers of software, making all the issues more broadly applicable. The urgency in addressing these issues will only grow in the coming decade.

[19] https://zenodo.org

6.4. *Community software as a force multiplier*

Collaboratively-developed community software has an increasing large impact throughout the astronomy community. For example, the whole scientific software ecosystem in Python (the most popular language in astronomy Momcheva & Tollerud 2015) is built on community-developed software like NumPy (van der Walt et al. 2011), SciPy (Jones et al. 2001–), Matplotlib (Hunter 2007), or other parts of the so-called "NumFOCUS Stack". More domain-specific projects such as Astropy project(Astropy Collaboration et al. 2013, 2018) and SunPy (SunPy Community et al. 2015) capture the expertise of a broad range of astronomers, and have a wealth of features that cannot be reproduced by solitary researchers. While the mere existence of such software open to all to use are immediately apparent, there are several ancillary benefits to such community software efforts:

- The more the community participates, the more the project will reflect their specific needs and applications, even if it is built on a more general framework.
- The code is typically inspected by more people, and many eyes make all bugs shallow (i.e. code problems and their solutions will be quickly found Raymond 2001).
- There is usually more documentation available because of the free energy to specialize on such tools, and a larger base to help support new users.
- It is easier to train scientists to help produce professional-quality software if they are supported by a core of professional engineers. Community projects provide a larger-scale social understanding of how that interaction can happen.
- These projects speed up the cycle of science by providing useful implementations for common tasks, freeing up researchers to work on their specific science.
- When built as part of an underlying broader ecosystem, community software often gains the direct benefit of contributions "upstream" e.g. improvements in core math libraries made by computer scientists can flow down to astronomy without any direct effort in astronomy.

Together, these factors mean that the impact of code developed by a community is multiplied by further contributions from other sources to the same ecosystem.

We note that the community developed software need not strictly be open source, though the majority of these projects are. The benefits of community development extend to both open and closed source projects, the primary difference being that the potential size of an open project is by definition larger than a closed one, and most of the above scale with community size.

6.4.1. *Open development/Open collaboration*

While a substantial fraction of software in Astronomy is now open source software, and has been for decades, a major development in recent years has been the growth of open development. This form of collaboration software development accepts and in many cases depends wholly on contributions from the wider community to the software project. Development of the code and discussion around that code is conducted in the open using industry-standard platforms like GitHub or GitLab, and in most cases policy discussions and decisions also occur in the open, or example on a public internet mailing list. The chief examples of projects like this in astronomy are The Astropy project and SunPy.

This kind of development model is not limited to astronomy projects, there are many examples of large scale software projects which are entirely developed in the open, the largest example of which is the Linux kernel. Developing software in this way introduces technical and sociological challenges, which have been met by DVCS tools such as git, online collaboration tools such as GitHub that enable workflows which scale to many hundreds or thousands of contributors, and the hard work of organizers and code reviewers to set up and maintain a positive culture that enables contributions to continue.

These kind of open collaborations enable many different stakeholders (both astronomer-users and dedicated developers) to collaborate on a software project, often from a diverse set of perspectives. While this is possible with non-open developed community software, it is often much harder because it requires an added layer of communication between "users" and "developers", while in open development these are the same community. This makes the software more valuable to both the contributors and the community more than the sum of the individual contributions, as it reflects the needs of the many rather than the one. It also means more work can be done with less funding, because the efforts of individual contributors are pooled into a "neutral" space that can arbitrate via the community process. Moreover, the open nature of the collaboration means that stakeholders have the ability to drive the direction and priorities of the project simply by contributing to it. Because many of these stakeholders are the users themselves, it also can serve to optimize the applicability-to-effort ratio.

6.5. *Community software problems and solutions*

With the above in mind, there is incongruity between the increasing importance of community software, and the funding available for such projects. In particular, the future of many widely used projects that are effective force-multipliers, including `astropy` and services such as astrometry.net, are uncertain. These major community projects are generally unfunded despite the vital role they play for astrophysics as a whole. While many feature "in-kind" contributions from user missions (as discussed above), such support depends on the vagaries of mission priorities rather than the needs of the community itself (as discussed below).

Hence, the benefits outlined above cannot be realized if such efforts are not supported by funding agencies, large missions, and indeed the astronomical community as a whole. Currently incentives are not in place to encourage community efforts: indeed in some cases such software development is either not allowed by a grant program, or tacked on as an afterthought. ("Oh, we'll probably have my grad student build that reduction pipeline on the way to their thesis.") Where software grant programs do exist, they often focus on building specific applications into interdisciplinary tools (e.g. NSF CSSI and DIBBs), rather than applying general software to specific domains. They also as a rule do not emphasize *community-building* elements like contribution policy documents, documentation of user workflows, or community coordination. Hence, while specific recommendations of what platforms for development are useful are not likely to be relevant in 10 years (and indeed are often counter-productive - see Section 6.6), our recommendations focus on incentives for pro-social behavior by missions and individuals. This will be critical to keeping up with the ever more software-rich Petabyte era, and this is precisely what the recommendations of this chapter aim to do.

6.6. *Software is alive*

> "This open source stuff is free. But it's free like a puppy. It takes years of care and feeding." - Scott Hanselman on the death of nDoc (Hanselman 2006)

The grant-funding model for academia fosters a picture of all work as limited to a fixed time horizon, shared astronomical software often lives as long as it is useful. This can be far longer than any individual researcher or developer, and as a result the software takes on a life of its own. Like any living thing, however, this software will not survive without proper care and feeding, and without evolving to adapt to continually changing environment.

6.6.1. *The software stack is always changing. We need to be adaptable.*

Sustainability is a necessary but not sufficient condition for software to survive. Even with maintenance, the entire software ecosystem is constantly evolving. A clear example is Python replacing IDL as the most popular programming language within astronomy (Momcheva & Tollerud 2015), despite many of the elements of the IDL Astronomy Library being maintained. Similarly, many of the features of IRAF are now being provided by widely used community projects such as Astropy project, despite the long history of IRAF. New software like this generally evolves because they can tackle problems that were not addressed previously, either by making the coding easier or taking advantage of other developments in the wider technical world (discussed more above). For example, resasons for the change from IDL and IRAF to Python are the lack of license fees, the extensive open source ecosystem of libraries for scientific computing, and the easier learning curve of the latter (due to more broad usage).

However, the disruption caused by the evolving software ecosystem can be disruptive because it comes at the cost of requiring significant retraining and refactoring. In this way, the need to be adaptable to changing developments in software can appear to be in tension with the need for well-validated software for research. There is indeed always a cost-benefit analysis for changing technologies that most include this concern as much as the benefits that may result. But consideration must be made that this disruption can be ameliorated by continuing education programs for researchers at all levels. Examples include AAS workshops, introducing astronomers to the up and coming software projects and to highlight long-term trends, such as which projects are growing in support vs which are now largely unmaintained. There are further more focused recommendations in Section 8 for keeping the community on top of such changes. Hence, the disruption caused by the continuous evolution of the software stack should not be feared, but rather welcome for its potential to improve our own research.

6.6.2. *Software needs to be sustainable*

Any software that is meant to be used more than once requires maintenance. Data sets change (or grow to Petabyte scale), bugs are discovered, computer architectures change, and users change their understanding of the intent of the software. This leads to the concept of software sustainability: practices both within the software itself and of those who develop it that make it practical to maintain the software for an arbitrarily long time. For astronomy software to be sustainable (Katz et al. 2018; Wilson et al. 2014), it should:

1. Be both testable and tested (i.e. it is correct and that correctness can be checked by anyone).
2. Be readable and useable by multiple people (i.e. it can evolve to fulfill its intent over time as development and scientific conditions change).
3. Have a viable pathway to be maintained past the original author (i.e. survives uncertainty).
4. Be able to respond to users' needs, even if they change over time (i.e. supports relevant concerns).

As outlined in Section 6.6.1, even for software that is maintained, for example by a third party organization (e.g. Harris Geospatial Solutions for IDL) does not guarantee future usage of this technology within astronomy (Momcheva & Tollerud 2015). As astronomy shifts towards a more community-developed, open source set of tools, it iss critical that different constituents of the astronomy community develop an understanding of the origin of this software and how they might be able to participate in its development, maintenance, and long term sustainability:

Software consumers (individual astronomers): Most individual researchers are *consumers* of community software, that is, they make heavy use of the software tools developed by their peers but do not routinely participate in the development of the software. Like most community-developed open source projects, this is the norm and is acceptable. However, complete ignorance of the origin of the software they are using creates a risk to the sustainability of the projects and individuals responsible for creating the software. For example, if they do not realize the software they are using comes from other researchers, they may not support hiring, tenure, etc of those who build that software, thereby stopping them from producing and maintaining the software itself. We believe therefore that even as software consumers, astronomers should increase their awareness of the origin of the software they are using and realize that they have an important role to play in the community by 1) providing feedback to software projects by filing bug reports, feature requests, feedback on existing tools, and perhaps contribute other resources like documentation if they have relevant expertise; 2) recognizing that software is created by *people*, and that supporting the work of their peers (be it financially, socially, or even emotionally) who spend time creating these tools is necessary for the tools they use to even exist; and 3) recognizing and advocating for the broader concept that using a shared set of community tools can improve all of science for less money.

Individual software creators (individual astronomers and engineers): While these are the bread-and-butter *of* these community efforts, they are not without shared responsibility here. Specifically, the builders of community have a responsibility for being aware of the community they are building for. E.g. they need to remember that the user community typically does not have as much technical expertise and therefore requires their *help* to both learn how to use the software and understand why it is useful. They also need to understand the unique responsibility that creating software sustainably is work (see the above subsections) and must either agree to such work or communicate clearly to their potential users that they cannot do it without help.

Institutional software creators (projects/missions/facilities): Observatories and missions (e.g. LSST, JWST, DKIST), especially in development & construction phases, spend significant resources developing software both for internal operations but also for their community to analyze and interpret data products from their facilities. These software creators need to be incentivized to *upstream* (i.e. contribute back new innovations to community software packages) their software where possible, thereby contributing to the large ecosystem of software available to the general astronomy community. As discussed earlier in Section 6.4, community software can be a force-multiplier when done right, but in order for this to happen, software projects must recognize their role in the community software ecosystem and shift towards being active contributors rather than consumers/users of community software.

7. ANALYSIS METHODS: ALGORITHMS AND STATISTICAL FOUNDATIONS

Contributors: *Brian Nord <nord@fnal.gov>, Andrew Connolly <ajc@astro.washington.edu>, Yusra AlSayyad, Jamie Kinney, Jeremy Kubica, Gautham Narayan, Joshua Peek, Chad Schafer, Erik Tollerud*

Note: If you have come directly to this chapter we suggest you please read at least the Introduction in Section 1 before delving further.

7.1. *Recommendations*

REC-13 Create funding models and programs to support the development of advanced algorithms and statistical methods specifically targeted to the astronomy domain.
Area: Analysis. **Audience:** Agency, Astronomer. **Term:** Medium
The increasingly large and complex datasets resulting from a new generation of telescopes, satellites, and experiments require the development of sophisticated and robust algorithms and methodologies. These techniques must have statistically rigorous underpinnings as well as being adaptable to changes in computer architectures.

REC-14 Build automated discovery engines.
Area: Analysis. **Audience:** Technologist, Astronomer. **Term:** Long
New hypotheses are difficult to generate in an era of large and complex datasets. Frameworks that can detect outliers or new patterns within our data could address many of the needs of current and planned science experiments. Funding and developing these engines as a community would lead to broad access to the tools needed for scientific exploration.

REC-15 Promote interdisciplinary collaboration between institutions, fields, and industry.
Area: Analysis. **Audience:** Agency, Manager, Astronomer. **Term:** Long
Expertise across multiple domains are required to tailor algorithmic solutions to astronomical challenges. The astronomical community should more heavily and directly engage researchers from industry and non-astronomy fields in the development and optimization of algorithms and statistical methods. Agencies and academic departments should develop funded programs to specifically connect astronomers to these experts through sabbatical programs, centers, fellowships, and workshops for long-term cross-domain embedding of experts.

REC-16 Develop an open educational curriculum and principles for workforce training in both algorithms and statistics.
Area: Analysis. **Audience:** Agency, Astronomer. **Term:** Medium
The speed of model and algorithm evolution requires regular training and education for scientists and for those seeking to enter science. Developing and maintaining open curric-

ula and materials would enable the teaching of algorithms and methodologies throughout the astronomical community.

REC-17 Encourage, support, and require open publication and distribution of algorithms.
Area: Analysis. **Audience:** Astronomer, Agency. **Term:** Short
The rapid adoption of advanced methodologies and the promotion of reproducible science would be significantly enhanced if we mandated the open publication and distribution of algorithms alongside papers.

7.2. *Overview*

The paradigms for data analysis, collaboration, and training have simultaneously reached a watershed moment in the context of algorithms and statistical methods. The onset of large datasets as a scientific norm accentuates this shift, bringing both technical opportunities and challenges. For example, the development of new algorithms and data modeling techniques has recently accelerated dramatically, providing new modalities for investigating large datasets. As this corner has turned in algorithmic development, the incorporation of rigorous statistical paradigms must keep apace. However, this shift has just begun, and we still lack the tools to even contend with, much less fully take advantage of, increasingly complex datasets for discovery.

The paradigm shifts also bring organizational challenges that highlight issues with cultural norms of education and collaboration about development of data analysis techniques. Discovery often occurs at the intersections of or in the interstices between domains, and therefore multi-dimensional collaboration has irrevocably become a key component of research. We need improved collaboration paradigms to take advantage of this accelerating emergence of technologies, thereby increasing the permeability of the barrier between different areas of science, and between academia and industry. Moreover, innovation in methods of education and training in new analysis techniques lag behind the development of the techniques themselves, leading to growing unequal distribution of knowledge. Similarly, accompanying software development strategies must keep apace with these developments, both to ensure results are robust and to make sure the education and training can be equitably distributed.

We have an opportunity to act as the changes set in and leverage our community's energy and inspiration to initiate change in how drive algorithmic discovery in the petabyte era. There is an opportunity for astronomy to both benefit from and help drive new advances in the emerging technologies. Below, we discuss the key challenge areas where we can and provide possible directions for what we can do.

7.3. *Discovery in the Petabyte Era*

At present the process of hypothesis generation in astronomy has two pathways. One is theoretical, wherein predictions from theory provide hypotheses that can be tested with observations. The other is observational, wherein surprising objects and trends are found serendipitously in data and later explored. As theory comes to depend on larger and larger simulations, and observational datasets grow into and beyond the petabyte scale, both of these pathways are coming under threat. With such large datasets, classical modes of exploration by a researcher are becoming prohibitively slow as a method to discover new patterns in data (e.g. finding objects and correlations by plotting up datasets). Without new hypotheses (and ways to develop them) in the 2020s, there may be no astronomy in the 2030s.

A key example of the challenge lies in explorations of high-dimensional datasets. Long ago, the discovery that stars fill an approximately 1D space in magnitude-color space led to a physical model of stellar structure. This is a low-dimensional, non-linear representations of higher-dimensional data. Indeed, seemingly smooth structures in astronomical data can have surprising substructure (e.g. the Jao/Gaia Gap (Jao et al. 2018)). 1D gaps in famous 2D spaces are visually discoverable. However, we lack comparable methods to find 2D gaps in 3D spaces, let alone structures in the extremely high-dimensional data that modern surveys create. Recently, Suzuki & Fukugita (2018) found 17 pure blackbody stars *by eye* amongst the 798,593 spectra in SDSS, nearly two decades after they were acquired. This result shows both how interesting outliers can be, and how by-eye methods are slow and not practical at the petabyte scale. With trillions of rows available in upcoming surveys, we'll have the ability to find low-dimensional substructure in high-dimensional that has potential to yield new physical insight — but only if we have the tools to do so.

As an example of such a tool, purpose-built Machine Learning (ML) algorithms coupled with deep sub-domain knowledge can successfully expose hitherto unknown objects that can significantly advance our understanding of our universe (e.g. Baron & Poznanski 2017). Unfortunately, any successful exploration requires a) deep algorithmic and implementation knowledge b) deep physical and observational domain knowledge and c) luck. Deep algorithmic knowledge is necessary as off-the-shelf algorithms usually need significant adaptation to work with heteroscedastic and censored astronomical data. Deep observational domain knowledge is needed as outlier objects are often artifacts and surprising trends may be imprints of the data collection method. Deep physical domain knowledge is needed to make sense of the result, and understand its place in the cosmos. For example, algorithms to find low-dimensional structures (McQueen et al. 2016, e.g. Manifold Learning;) are only one piece. Observational expertise is necessary to determine that the observed manifolds are real, and astrophysical expertise is necessary to formulate physical explanations for the observations. Finally, not all searches will return results; a modicum of luck is needed. This trifecta of algorithmic knowledge, domain knowledge, and luck is rare.

Over the next decade, we expect astronomy to require unique, fundamental new developments in algorithms, statistics, and machine learning. Despite the incredible pace of innovation within these fields, it will not be enough for astronomy to ride along and adopt general technologies. Astronomy's science drivers will bring unique algorithmic and statistical questions, data characteristics, and edge cases that will both require and drive continued investment and innovation.

We argue that the path forward is through the construction of intuitive, trustworthy, robust, and deployable algorithms that are intentionally designed for the exploration of large, high-dimensional datasets in astronomy. When we consider the current landscape of astronomical research and the upcoming generation of sky surveys, we can already identify areas where algorithmic and statistical investment are needed, such as:

1. Online (i.e. close to real-time) alerts and anomaly detection in large sky surveys will require high throughput algorithms and models in order to keep up with the volume of data produced.
2. Statistical and learned models need to go beyond black box optimization. Models should be understandable and interpretable in terms of the physical systems they represent.
3. Machine learning algorithms may need to be adapted to make effective use of domain knowledge such as physical constraints and data collection methodology.
4. Machine learning techniques often introduce new parameters that must be recorded in a standardized form to allow other researchers to reproduce analysis.

Very few researchers have both all the needed skills and the bravery/foolhardiness to seek out risky avenues of research like these. We therefore propose that funding agencies fund the creation and maintenance of “discovery engines” — tools that allow astronomers without deep algorithmic knowledge to explore the edges of data spaces to hunt for outliers and new trends. These engines should be hosted near the data when needed (Section 5.2), but should be initiated by the astronomical and methods-development communities.

The development of new statistics and algorithms can be accomplished through a variety of methods, including: on-boarding dedicated algorithmic/data-intensive science experts onto astronomy teams, facilitating partnerships (with industry or other academic fields), and building internal expertise within the community through education and training. Regardless of the mechanism, it is important that the development of new statistical and algorithmic techniques is considered a core part of astronomical missions.

7.4. *The state of statistics: statistical methodologies for astrophysics*

Statistical methods and principles are the backbone upon which successful estimation, discovery, and classification tasks are constructed. The tools commonly associated with Machine Learning (e.g. deep learning) are typically efficient, “ready-to-use" algorithms (albeit with ample tuning parameters). On the other hand, statistical approaches employ a set of data analysis principles. For example, Bayesian and frequentist inference are two competing philosophical approaches to parameter estimation, but neither prescribes the use of a particular algorithm. Instead, the value (and perhaps the curse) of the statistical approach is that methodological choices can be tailored to the nuances and complexities of the problem at hand. Hence, when considering the statistical tools that are crucial for astronomy in the coming decade, one must think of the recurring challenges that are faced in data analysis tasks in this field.

Further, as the sizes of astronomical survey datasets grow, it is not sufficient to merely "scale up" previously-utilized statistical analysis methods. More precisely, modern data are not only greater in volume, but are richer in type and resolution. As the size and richness of datasets increase, new scientific opportunities arise for modeling known phenomena in greater detail, and for discovering new (often rare) phenomena. But these larger datasets present challenges that go beyond greater computational demands: they are often of a different character due to the growing richness, which necessitates new analysis methods and therefore different statistical approaches. Hence, as the complexity of astronomical data analysis challenges grow, it is imperative that there be increasing involvement from experts in the application of statistical approaches.

To ground these ideas, in the following subsections we will consider examples of technical and organizational challenges that, if advanced over the next decade, would provide the greatest scientific benefit to astronomy.

7.4.1. *Technical Challenges*

1. *Methods for the analysis of noisy, irregularly-spaced time series.* Future time domain surveys, like LSST, will generate a massive number of light curves (time series) with irregular observational patterns and in multiple bands. This goes beyond the limits of classic time series models, which assume regularly spaced observations with a simple error structure. Areas of need include feature selection for classification, periodicity detection, and autoregressive modeling,
2. *Likelihood-free approaches to inference.* Likelihood-based inference is standard in astronomy, but as the sizes of datasets grows, any flaw in the assumed likelihood function will result in a bias in the resulting inference. Such flaws result from unwarranted Gaussianity assumptions, difficult-to-model observational effects, and oversimplified assumptions regarding measurement errors. Likelihood-free approaches, such as approximate Bayesian computation, hold promise in astronomy, but much work is required to develop tools and optimize them for astronomy datasets and therefore make this computationally-intensive approach feasible.
3. *Efficient methods of posterior approximation.* Even in cases where a likelihood function is available, constructing the Bayesian posterior is challenging in complex cosmological parameter estimation problems, because future inference problems will push the computational boundaries of current MCMC samplers. Work is needed to improve the performance of chains, which must adjust to degeneracies between cosmological parameters, handle a large number of nuisance parameters, and adhere to complex hierarchical structure that is increasingly utilized in such analyses.
4. *Emulators for complex simulation models.* It is increasingly the case that a simulation model provides the best understanding of the relationship between unknown parameters of interest and the observable data. Unfortunately, these simulation models are often of sufficient complexity that a limited number of simulation runs can be performed; the output for additional input parameter values must be approximated using emulators that interpolate these available runs. Emulation to sufficient accuracy

requires careful selection of both the input parameters for the training sample and the method of interpolation; both of these must be done with consideration of the particular application.

5. *Accurate quantification of uncertainty.* Complex inference problems in astronomy are often, out of necessity, divided into a sequence of component steps. For example, classification of Type Ia supernovae, a challenging problem on its own, is just a step in a larger analysis that seeks to constrain cosmological parameters. Separately, redshifts and luminosity functions are estimated and then fed into larger estimation problems. This divide-and-conquer approach requires careful consideration of the propagation of error through the steps. How does one quantify errors in redshift estimates in such a way that these uncertainties are accurately accounted for in the downstream analyses? How is contamination that results from misclassification of supernovae reflected in the uncertainties in cosmological parameters estimated from these samples? LSST faces challenges of separating identifying images in which overlapping objects are “blended"; how is the uncertainty inherent in this problem incorporated into analyses that use these images? Careful consideration of such questions is crucial for attaching accurate statements of uncertainty to final estimates.

7.4.2. *Organizational Challenges*

1. *Accessible publishing of methods.* Advances in statistical theory and methods abound in the literature of that field, but it is often presented in a highly formalized mathematical manner, which obscures the aspects of most importance to potential users. This creates a barrier to the appropriate use of these methods in astronomy. The greater involvement of data scientists in collaborations will help to bridge this divide, and enable these individuals to make significant contributions. This will require appropriate professional recognition for this effort, including encouraging the publication of methodology papers in astronomical journals by data scientists (see Section 8.5.2 for related workforce issues).
2. *Avoiding the “algorithm trap."* Astronomical inference problems are of sufficient complexity that full use of the data requires analysis methods to be adapted and tailored to the specific problem. For this reason, statisticians prefer to not think of an analysis as the application of a ready-made “algorithm.” By contrast, astronomers are generally more interested in the result of the analysis, so are attracted to well-separated “algorithms” they can apply to a problem. This difference in perspective only increases the need to have data scientists deeply involved in the collaborative process.
3. *Reducing barriers for statisticians.* From the other side, data scientists face challenges in applying analysis techniques astronomical data. This is partly due to technical difficulties like unique file formats and data access issues. But it is also because deeply understanding the science is frequently crucial to building methods tailored to the problem, as outlined above. More effort needs to be placed on reducing these barriers. For example, astronomers can work to isolate important statistical

aspects of larger problems and create user-friendly descriptions and datasets to allow statisticians to more quickly learn and focus on making a contribution. At the same time, embedding statisticians and data scientists close to astronomers will help bring the former to a better understanding of the astronomy perspective.

7.5. *The state of algorithms: developments in the last decade*

A key development that has enabled science in this past decade has been the development of a number of general purpose algorithms that can be applied to a variety of problems. These algorithms, irrespective of what programming language the implementation is in, have made astrophysical research more repeatable and reproducible, and less dependent on human tuning.

For example, PSF kernel convolution has enabled time-domain astrophysics, and is a key component of difference imaging pipelines, but is also used to generate deep stacks of the static sky, allowing us to find ever more distant galaxies. These developments in turn have spurred the development of new algorithms. In roughly 20 years, the field has moved from Phillip Massey's guide to doing aperture photometry by hand with IRAF for small, classically scheduled programs, to completely automated surveys that optimize their observing schedule in real-time, record data, detrend the observations, and perform automated PSF photometry of billions of deblended sources.

As with statistics, the distinction between algorithms, and the software implementation of algorithms is blurry within the community. In many situations, we now use algorithms without any knowledge of how they work For example, we can now expect to sort tables with millions of rows on multiple keys, without knowing the details of sorting algorithms, precisely because these details have been abstracted away. We note that the many widely used algorithms, such as affine-invariant Markov Chain Monte Carlo techniques are widely used precisely because the algorithm is implemented as a convenient software package. Community-developed software packages such as `scikit-learn`, `astropy`, and the IDL Astronomy Library have increased the community's exposure to various algorithms, and the documentation of these packages has in many cases supplanted implementation-oriented resources such as Numerical Recipes.

At the same time in the broader world, a class of algorithms is being used to execute tasks for which an explicit statistical forward model is too complex to develop, and correlations within the data itself is used to generate actionable predictions. These AI techniques include machine learning models, which have been used to replace humans for tasks as varied as identifying artifacts in difference images, to categorizing proposals for time allocation committees. These AI techniques, in particular deep learning methods, are increasingly viewed as a solution to specific petabyte scale problems, as they have been successfully deployed in the commercial sector on these scales. We anticipate increasing adoption of these algorithms, as user friendly implementations such as `pyTorch` and `Keras` become more well known, and data volumes grow. It is also likely that the algorithms that are used to train these machine learning methods, including techniques like stochastic gradient descent, will find more use within the astronomical community.

Machine learning algorithms are necessary but not sufficient to continue the progress in astrophysical research that is driven by algorithms. In particular, machine learning methods are often not-interpretable, and while their output can be used effectively, those outputs are not true probabilities. The scientific method fundamentally involves the generation of a testable hypothesis that can be evaluated given data, and is therefore inherently statistical. As data volumes grow, the dimensionality of models grows, and there is increasing recognition that the model structure is hierarchical or multi-level. While we see increasing adoption of hierarchical models for Bayesian inference, there remains much to do to increase awareness of algorithms to effectively evaluate these models, including probabilistic programming - algorithms that are used to build and evaluate statistical models in a programmatic manner.

As in the previous section, we now separately consider some of the specific technical and organizational challenges in the area of algorithms.

7.5.1. *Technical challenges*

1. Both algorithms and models need to be trustworthy and interpretable. It's easy to throw a dataset into a neural net or ensemble classifier and overfit. Tools need to be developed that recognize these traps and in large-scale datasets, and bring them to the attention of the user.
2. Many algorithms, especially in the machine learning space, require labeled data that may not be available at sufficient volumes, or at all.
3. The reproducibility of results derived from algorithms needs to be improved. This is especially important with machine learning models where black-box optimization is often used because it is an easy-to-provide feature. Such reproducibility improvements could be as simple as defining standardized formats for how we document the model learning parameters, but could also be more complex, including building out tools that are designed specifically for reproducibility (e.g. Data reduction pipelines with built-in provenance, or Jupyter notebooks that download their own data).
4. Scalability of newly-developed algorithms. With the data volumes of the petabyte era, efficiency in all parts of the stack is necessary. Such optimizations are usually possible, but require investment of time (often by different people than those who develop the first iterations of the algorithm).
5. Astronomy data has some differences that can expand current algorithmic development at large. This particularly includes use of measurement uncertainties, as general-use algorithms often make assumptions that work for other fields that are homoscedastic or Gaussian which fail in Astronomy. There is also a need for more algorithms that account for posteriors, a particularly strong need in astronomy because its domain of "the universe as a whole" means that algorithms applied to one dataset need their outputs to be considered by another.

6. Significant work is still needed in adapting and improving the current space of existing algorithms: optimizing traditional astronomy algorithms, adapting them for a cloud setting, or even making small accuracy improvements.

7.5.2. *Organizational challenges*

1. It is difficult to get the necessary expertise onto all missions that will need it both in terms of developing the expertise internally (due to the fast pace of change in the space) and hiring in experts.
2. There is currently no established marketplace/mechanism for matching difficult problems in the astronomy domain to relevant experts outside an astronomer's network. This is particularly acute given the discussion above about the growing importance of statistical and data science expertise.
3. There is a missing component in the conduit of moving new algorithms developed in academia into robust, usable, finished products. See Section 6 for additional discussion in this area.
4. We need standardized processes for publishing algorithms and machine learning models such that the results obtained with these algorithms/models are: broadly accessible, discoverable, fully reproducible (including archiving the model parameters), and easily comparable with other algorithms in the problem space.
5. We need to define and fund a process for continually modernizing/upgrading algorithms as the broader environment changes (new languages, new libraries, new computational architectures, shift to cloud computing, etc). See Section 6.6 for a broader discussion of mechanisms and recommendations for this.

7.6. *Emerging trends in industry and other fields*

Over the past two decades, the wider industry has also seen a shift in development approaches and computational techniques that can be adopted by the astronomical community. As noted in Section 6 open source software has become a new normal with communities sharing their investment in software development. When considered along with the industry's shift toward cloud computing and software as a service, astronomy can benefit from the new scale and availability of off-the-shelf solutions for computation and storage. Astronomers no longer need to focus significant portions of time on the low-level technical details in running dedicated banks of computers to support each survey.

This service model is being extended beyond software deployments and starting to push into algorithms as a service. Cloud machine learning services provide a portfolio of general algorithms. Instead of worrying about the specifics of the algorithm development, users focus only on model specification. This requires a shift in how we think about new algorithm development. Instead of focusing on the details such as implementation, optimization, and numerical accuracy, the practitioner focuses primarily on the high level model specification. Due to a series of recent successes, a significant focus within hosted machine learning services has been on deep neural networks (DNNs). NNs have shown remarkable success across a variety of tasks. Further new developments such as convolutional neural networks and recurrent neural networks have extended the power of this technique.

Another area of focus within the field of machine learning is blackbox optimization. Techniques such as Gaussian decision processes, allow algorithms to jointly model and optimize unknown functions. These techniques can be applied to a range of problems from optimizing real-world, physical processes to optimizing the parameters of a machine learning system (e.g. AutoML).

The ultimate goal of algorithms as a service can be seen in the advancements in AutoML. AutoML systems aim to abstract away not just the algorithm's implementation details, but also the need to manually tune model parameters. For example, recent work in Neural Architecture Search (NAS), allows the AutoML system to handle such development decisions as choosing the structure of the network (number and width of layers) as well as the learning parameters. While this automation greatly simplifies the problem of constructing accurate models, it does move the practitioner one step further from understanding the full details of the model.

There is an opportunity for astronomy to both benefit from and help drive new advances in the emerging industries. As noted above, astronomy can benefit from the shift from individually developed and maintained systems to hosted platforms that allow more effort to be spent on the data analysis itself. Moreover, the shape and size of science data serve as a driver for the development of new algorithms and approaches. We expect many of the upcoming advancements to be driven by real-world problems—machine learning will rise to the challenge of solving new, open problems. The recommendations in this chapter aim to ensure some of these problems and solutions are in the astornomy domain.

7.7. *Enhancing Interdisciplinary Programs and Collaborations*

The past decade has been a period of rapid change in the the multi-dimensional landscape of algorithms, computing, and statistics. We have seen the rise of new "standard" programming languages and libraries (e.g. Python, `astropy`, `scikit-learn`). There has been a proliferation of new algorithmic and statistical techniques — from improvements in image processing and compression to the rise of deep neural networks as a powerful tool from machine learning. We have seen the rise of new computational modalities, such as cloud computing and software as a service. New distributed compute frameworks such as Dask and Spark are emerging to process and analyze large and complex datasets. Even the basic mechanics of computation is undergoing a shift with the availability of specialized hardware such as GPUs and TPUs, requiring a new domain of knowledge to efficiently deploy solutions. There is no reason to expect the pace of innovation to drop off anytime soon.

This rapid pace of advancement means that it is no longer possible for a single astronomer or even a small team of astronomers to build the necessary depth of expertise in all of these areas. However, these technologies are already proving critical for maximizing the scientific reach of new research. Robust methodologies that can scale to the expected size and complexity of the data from new astronomical surveys and experiments will need to be accessible and usable by a broad section of our community. As new technologies spring up quickly, the astronomical community will need to balance the cost of learning the new technologies with the benefits they provide. It is not reasonable to expect every astronomer to keep up with all of the advances. A number of new ad hoc collaborations or collectives have sprung up to bring together astrophysicists and deep learning experts, such as the Deep Skies Lab[20] and Dark Machines[21].

In cases where collaborations exist today, there can be a variety of complicating challenges. There is currently no established marketplace for matching difficult problems in the astronomy domain to relevant experts outside an astronomer's network (see also §7.6). The resulting in-depth collaborations have start up overhead as the external experts learn enough about the problem domain to be helpful. Short-term engagements can suffer from a lack of depth or insufficiently productionized solutions. Even in longer term engagements, there can be misalignment between the parties due to the different incentives. For example, statisticians and computer scientists in academia are primarily recognized for only the novel contributions to their own fields. Papers that apply existing methodologies to new problems are not considered significant contributions to their fields. Similarly, members of the astronomy community are not fully recognized for their algorithmic contributions.

There are many opportunities for astrophysics to benefit from these investments in technology and computational algorithms. However, requires that we change how astronomy engages with experts in other fields. The exact shape of this engagement can take a variety of forms. Examples include:

1. Provide funding for astronomical missions to engage with external experts (academic or industrial) via consulting, co-funded research, or subcontracting.
2. Encourage a robust community of volunteers via open source contributions and engagement.
3. Create forums for external methodological experts to engage in astronomical projects and analyses. Data challenges and hack sessions can be used to encourage engagement, but they require sufficient organization and communication (i.e. funded effort) to ensure they can engage software engineers at an appropriate level.
4. Encourage recognition of interdisciplinary contributions within academic areas (e.g. career progression for statisticians that enable new astronomy without necessarily creating new statistics).
5. Organize workshops that bring together members of these different fields and can facilitate matching along problem domain.
6. Provide funding for astronomical programs to hire full time experts to be embedded within the mission. It is important to note that this approach comes with challenges in recruiting (both these areas are in high demand), costs of attracting high quality

[20] deepskieslab.ai
[21] darkmachines.org

personnel, and in stability for the team members (the algorithmic / statistical workload might not be consistent throughout the life of a project).

7. Implement reverse sabbaticals where experts from industry can embed in projects for short intervals (a few months).
8. Train astronomers in these fields to become resident experts. Encourage mobility of these experts to provide support for new missions.
9. Establish a center for algorithmic and statistical development in astronomy (centralized or virtual) that employs full time experts in fields such as algorithms, statistics, and machine learning. This center would be a community resource that provides support to individual programs via deep engagement.

The goals of these interactions are not to provide programming support for projects but to develop a base of expertise built from academic and industrial experts that can help to define, design, and guide the development of computational and statistical projects within astronomy. The form and depth of the engagement will naturally be project dependent. Experimental and privately-funded interdisciplinary centers e.g. the Moore-Sloan Data Science Environments at Berkeley, New York University (NYU) and the University of Washington, or the Simon's Flatiron Institute have demonstrated how expertise in data science can advance a broad range of scientific fields. Access to the resources at these centers is, however, limited to researchers at these privileged institutions. The challenge we face is how to scale these approaches to benefit our community as a whole.

7.8. *Education and training*

Training a workforce that can address the algorithmic and statistical challenges described in this Chapter will require a significant change in how we educate and train everyone in our field, from undergraduate students to Principle Investigator (PI)'s. The discussion in this section is complementary to and aligned with that found in Section 8 and Section 9. The traditional curricula of physics and astronomy departments do not map easily to the skills and methodologies that are required for complex and/or data intensive datasets. This is a rapidly changing field, and will remain so for at least a decade. However, a strong foundation in Bayesian statistics, data structures, sampling methodologies, and software design principles would enable professional astronomers to take advantage to big data in the next decade. Bridging this gap between the skills we provide our workforce today and the ones they might need to succeed in the next decade should be a priority for the field.

In the previous decade there was substantial progress in creating material to support the teaching of statistics and machine learning in astronomy. This includes the publication of introductory textbooks (Ivezić et al. 2014; Kohl 2015; Hornik 2018), the creation of common software tools and environments (Astropy Collaboration et al. 2013), the development of tutorials, and a growing focus on software documentation (Astropy Collaboration 2019). The emergence of Jupyter (Kluyver et al. 2016) as a platform for publishing interactive tutorials and Github and Gitlab for hosting these tutorials and associated code has simplified the process of sharing material. To date, however, there has been little coordination in this effort. The coverage of topics in the available material is not uniform. Moreover, the underlying principles and foundations of statistics are often not covered in favor of the introduction of commonly used software tools and algorithms. For the case of algorithmic design and optimization there has been substantially less progress in training the community. Instead, the primary focus being the development of introductory materials such as the Software and Data carpentry (von Hardenberg et al. 2019; Wilson 2013).

We have started to make progress in providing an educational foundation in statistics and algorithms, but it is not uniformly available across our community — with significantly less access at smaller colleges and in underrepresented communities. We, therefore, recommend the development and support of a common and open set of educational resources that can be used in teaching statistics, and algorithms, and machine or computational learning. Determining what constitutes an appropriate curriculum will be a balance between providing the foundations of statistics and algorithmic design appropriate for the broader science community and teaching specialized skills (e.g. optimization, compilers) that may benefit a smaller, but crucial, set of researchers who will engage in the development and implementation of computing and software frameworks.

This will likely require a coordinated effort to integrate current resources within a broader curriculum and to make them easily accessible — in a manner where anyone, from astronomer to an entire educational institution, can create custom courses tailored to their needs. Given the rapid evolution in algorithms and in the ecosystem of tools over the last decade, and looking to the future, this curriculum will need to be able to evolve.

8. WORKFORCE & CAREER DEVELOPMENT

Contributors: *Dara Norman <dnorman@noao.edu>, Kelle Cruz, Vandana Desai, Britt Lundgren, Eric Bellm, Frossie Economou, Arfon Smith, Amanda Bauer, Brian Nord, Chad Schafer, Gautham Narayan, Ting Li, Erik Tollerud*

Note: If you have come directly to this chapter we suggest you please read at least the Introduction in Section 1 before delving further.

8.1. *The growing importance of a tech-savvy workforce of astronomers*

In the rapidly approaching era of large surveys, experiments, and datasets, we will only reach our scientific goals if we train and retain highly capable scientists, who are also engaged with technological advances in computing. With the goal of advancing scientific discovery through the collection and analysis of data, we must commit and dedicate resources to building both the skills and competencies of this workforce. This includes those in the workforce that will be using data to advance science as well as, those supporting the infrastructure that make those discoveries possible. The areas and skill sets in which we our teams need training are software carpentry, algorithms, statistics, the use of tools and services (for scientific staff); and software engineering effective practices, data management and access (for support staff).

In this chapter we discuss the activities needed to build, support, and advance the scientific workforce that will take the petabytes of data collected to scientific discoveries over the next decade. In particular, Section 8.2 discusses the current demographics of the data science support mission, exemplifies the scope of training (Section 8.3) that is needed to build this workforce. Section 8.4 focuses on training for researchers who are more accurately described as "users." In Section 8.5, we discuss modern challenges for these career paths, as well as how to address them. Finally, in Section 8.6, we identify metrics that we should be using for training in career development and for reviews in career advancement.

8.2. *Demographics - who is this workforce and where are they now*

Data support roles permeate the astronomy and astrophysics (hereafter, "astronomy") science community, and they encompass people with a variety of job types and descriptions and at levels from post-baccalaureate to PhD. A range of experience with either topics of astronomy or computing also differentiate roles. This range of data support positions requires a diversity of opportunities for training to work at the various levels, as well as career development and advancement suited to those career tracks. For example, positions for those with PhDs are significantly different from those that require only a post-bac degree, and thus the metrics used to support and determine career advancement must also be different. It has only recently been recognized that this role should be trained for and tracked independently of scientific interests and other professional duties. Consequently, the community has not adequately tracked the quantity and demographics of astronomy researchers currently engaged in science data support roles.

Instead of quoting statistics here, we present exemplar descriptions of current job titles and roles. Many of the people engaged in science data support hold PhDs in astronomy, astrophysics or physics. These researchers may be employed at colleges, universities, data centers or observatories, national laboratories. They may hold a leveled academic title (e.g. Professor, Astronomer, Scientist, etc.), as well as an additional functional job position in centers or programs with names "Data Science Mission Office," "Community Science and Data Center," "Infrared Processing and Analysis Center." Meeting career milestones to move up the ladder in these academic titles (i.e. assistant, associate, full, etc.) currently often only include the same metrics as for other faculty and staff (e.g. numbers of published papers, h-value, etc.) More discussion is in Section 8.5.

There are also many other science data support roles, in which staff have degrees at the BS, MS, or PhD level with position titles like "research and instrument associate," "research and instrument scientist," "mission systems scientist," "archive scientist." These staff are often responsible for coding and database support. Below, we discuss the resources and cultural changes needed to support the career trajectories of this workforce, to slow the threat of "brain drain" from the field, and to develop a workforce that can thrive in academia, industry, or government lab positions.

8.3. *Training to contribute to software development: Building the next generation*

Astronomers have a long history of developing useful software, but software development itself has not been considered a core component of the astronomy training curriculum. The expectation of petascale datasets in the 2020's provides a strong motivation to increase familiarity with effective practices in software development, as well as with existing frameworks that are widely used in the commercial sector. This cultural change will lead to better software in astronomy and more innovative scientific discovery. It will also provide astronomers with invaluable training that will increase their familiarity with (and marketability to) work in industry.

Currently, effective practices include using version control (e.g. GitHub), maintaining documentation and unit tests with code, and employing continuous integration methodologies, in which code is built and executed in shared repositories, allowing teams to identify issues early. Analysis in the 2020s will involve many pieces of software that are integrated into complex pipelines, processing ever-larger volumes of data. Astronomical projects are now comparable in scale to large industrial software development projects. Consequently, the gap between these effective practices and the modern cultural norm in astronomy and astrophysics must be reduced as the field transitions to increasingly large collaborations.

The increasingly critical role of software development in astronomy clearly indicates it is crucial that software development become part of the core graduate curriculum alongside typical coursework, like mathematics and observing techniques. Such coursework will also help reduce the disparity between students from diverse backgrounds, some of whom may never have been exposed to software development, or even coding, as undergraduates. This course material is distinct from, but complements training in data science and scientific computing techniques, which are increasingly being incorporated into Astronomy coursework. Developing the course material for data science work is likely beyond the scope of most departments, but vital steps have already been taken by several groups. Notably, the LSST Data Science Fellowship Program has already developed materials to expose students to best practices for software development. Curating these materials, and augmenting them with information on widely-used platforms will reduce the barrier to adopting such coursework or integrating it into existing classes.

There are several other challenges for supporting scientific software training in a university setting. One challenge is lack of access to state-of-the-art technologies: the landscape of coding and software development changes rapidly as coding languages come and go, workflow best practices continually evolve, and new platforms emerge and gain wide acceptance. For principal investigators and project managers to make informed decisions and guide their teams, there must be opportunities for them to stay abreast of these developments and to evaluate their utility even if they are not the ones actually using the various tools.

Another challenge resides in the structure and processes of university departments. Many computer science departments do not teach the programming skills necessary for scientists. Thus, the burden of developing more appropriate materials is fractured and currently falls upon individual instructors. The field needs dedicated staffing to develop curriculum materials for computational training. A fundamental barrier to the development of reliable, curated, and widely shared software in astronomy is the lack of incentives for this work and the dominance of the "publish or perish" mentality. Changing this cultural norm requires that our community incentivize — both within scientific projects and across the field at the employment level — work in developing good software and in educating people to build good software. Recognizing such work in assessing service and research, and valuing well-written and -documented software that is widely used for scientific work, rather than only immediate personal results is a vital step in changing this culture and in preparing the field for the software challenges that will be posed by massive projects in the 2020s. A full solution cannot be realized through universities alone, and partnerships with data centers, observatories, national labs, and professional societies are crucial.

The clear successes and popularity of the various existing training programs, which grew organically out of the community, attest to the need for additional and more advanced training resources. While there are several successful programs that address some of these concerns, they are insufficient to meet the needs of the larger community. For example, the Software Carpentry curriculum (https://software-carpentry.org/lessons/) is limited to the very basics of version control and collaborative software development but does not cover topics, like performance optimization, continuous integration and testing, and documentation. Furthermore, most of these workshops are targeted to senior graduate students, with a few targeting very early-career scientists, and they are not designed to meet the needs or concerns of mid-career scientists and managers. Thus, these programs are currently limited to a very small portion of the community and are currently unable to provide the needed training to people in multiple sectors of our community who need and want these opportunities.

Staff at Data Centers may themselves currently lack data science skills and up-to-date knowledge. Funding to support career development for current staff and to enable centers to hire staff that have data science expertise is critical to building workforce capacity in the 2020s.

Fundamental coding and software development skills are becoming increasingly necessary for success in every aspect of Astronomy. However, acquiring professional training in these skills is rare and inaccessible or impractical for many members of our community. Students and professionals alike have been expected to learn these skills on their own, outside of their formal classroom curriculum or work duties. Despite the recognized importance of these skills, there is little opportunity to learn and build them — even for interested researchers. To have a workforce capable of taking advantage of the computational resources and data coming in the next decade, we must find and support ways to make coding and software development training widely accessible to community members at all levels.

REC-18 Programs to cultivate the next generation.
Area: Workforce. **Audience:** Agency . **Term:** Long
Agencies should fund more and large-scale programs that cultivate the next generation of researchers versed in both astrophysics and data science, similar to smaller and over-subscribed programs like Software and Data Carpentry, LSSTC Data Science Fellowship/ La Serena Data School for Science, Penn State Summer School in Statistics for Astronomers.

REC-19 Support to produce training materials.
Area: Workforce. **Audience:** Agency . **Term:** Short
Provide funding to data and computational centers to produce modular and re-usable training resources to the community. These resources should be designed to be used by individuals, integrated into formal classes, and used as part of professional development training.

REC-20 Long-term curation of materials.
Area: Workforce. **Audience:** Agency . **Term:** Long
Funding must be provided to host and support educational materials in a long-term, stable, scalable place. Provides stability and improves discoverability if materials can live in a centralized location.

REC-21 Funding for innovative partnerships.
Area: Workforce. **Audience:** Agency . **Term:** Medium
Incentives should be provided to launch opportunities to harness partnerships between data centers, universities and industry through funding. For example, support for sabbatical

programs at the data centers where teaching faculty can learn skills, develop educational materials for community use, and bring back to their home institutions.

REC-22 Software training as part of science curriculum.
Area: Workforce. **Audience:** Astronomer, Educator, University . **Term:** Medium
Individuals, departments, and professional societies should encourage educational programs to incorporate software training skills into their existing courses and programs.

8.4. *Training to take advantage of big data for research: Bytes to Science*

Astronomers who came of age before the era of Big Data require training to take advantage of astronomical "Big Data" in the 2020s. They also need these skills to mentor students, who are simultaneously learning both astrophysics and the uses of data for research. It is crucial that access to this training be made widely available to professionals who come from a variety of science backgrounds and are based at a broad range of institutions (e.g. universities, data centers, etc.). This is especially important, considering these professionals will be cultivating their students and the next generation of scientists, as well as making decisions about which technologies to invest in. If access to advancing data skills remains difficult to obtain, we will fail to build a diverse workforce equipped to answer the most pressing questions in astronomical research. Data Centers could play an important role in providing this training.

New, freely accessible open source code and Jupyter frameworks like SciServer.org and NOAO Data Lab enable anyone with a web browser to quickly and easily analyze vast stores of professional astronomy datasets via web-based notebooks. These cloud-based platforms can democratize educational access by providing a scale of computing power and data storage that was previously reserved for students and faculty at well-resourced research institutions, where high-performance computing access and support are abundant. A small number of astronomers in higher education are already developing instructional activities for these platforms. These instructional materials train students and other users to explore and analyze large professional astronomy datasets with ease and to equip users with the computational foundation needed to pursue advanced independent research projects.

Jupyter notebooks in particular hold enormous potential for training the current and next generation of astronomy professionals. However, currently, the development of standardized curricular activities is performed in an entirely ad-hoc manner. Limited resources (funding and time) lead to very little deliberate coordination amongst various astronomy faculty who produce such materials, and these products are not sufficiently discoverable (e.g. accessible through a common repository).

The establishment of Community Science Centers hosted by Data Centers (like NOAO) can be a hub (clearing house) to bring information to the community about opportunities for the kind of resources and training that allow a broad group of researchers to go from petabytes to publications.

In order to provide the most useful training, data centers need a clear view of user needs. This information is provided by advisory committees, like "User Panels." However, these panels are traditionally populated by astronomers based at R1 institutions and other data centers. Data Centers should ensure that their User Panels include representatives from small and under-resourced institutions; this will provide a clearer picture of the unique training needs and challenges that must be addressed for these researchers. In addition community surveys that reach astronomers who do not currently use data centers should be undertaken to better understand what barriers exist.

REC-23 Training activities and materials.
Area: Workforce. **Audience:** Agency . **Term:** Short
Agencies must ensure that big data science training activities and materials for PROFESSIONALS (as well as students) are included as part of the federally funded data center's mission and deliverables.

REC-24 Change advisory board representation.
Area: Workforce. **Audience:** Agency . **Term:** Medium
Federally (and privately?) funded science centers should include representatives from small and under-resourced institutions to provide a broad and clear picture of need in the community. The collection of information, perhaps through surveys, to better understand the barriers to access that exist for astronomers at these institutions should be undertaken by data centers and others.

8.5. *Identifying career paths around scientific software development & big data science*

The key skills necessary for data-intensive scientific research are also highly valued in industry, government, and media/communication sectors. Astronomy training can serve as a stepping stone to fulfilling careers in a wide variety of fields, and astronomers should support and encourage those who transition to jobs in non-academic science, because ties with industry can strengthen and leverage our partnership opportunities. However, we need informed people on both sides: in many cases, challenging and uncertain career paths in astronomy push the best and brightest towards careers where their contributions are more readily appreciated. This "brain drain" siphons away the very researchers most needed to tackle the most pressing science questions of the 2020s.

8.5.1. *Universities*

In the university context, tenure-track faculty positions remain the gold standard for stability, compensation, and prestige. However, despite the fundamental role of software in scientific discovery, it remains difficult to receive credit towards tenure and promotion for developing software and services. Section 8.6 offers more specific recommendations for improving recognition for these contributions.

Even with appropriate credit for software contributions, faculty positions will continue to carry expectations of leadership, grant-writing, teaching, mentorship, and service, as is appropriate. Furthermore, driven by ongoing changes in the landscape of higher education, tenure-track hiring continues to flatten. To benefit from the opportunities of large datasets, universities also need the ability to support and retain technically capable faculty and staff, who have expertise and a longevity that typically cannot be matched by graduate students or postdocs. These “Research Software Engineers” (Research Software Engineers International (2018)) would provide a technical core for data-intensive research groups, just as opto-mechanical and electrical engineers are vital to the success of instrumentation labs.

Stable funding is the largest need for the success of staff Research Software Engineers (Geiger et al. 2018). A patchwork of 2-3-year soft-money grants is insufficient to retain highly-capable professionals, especially when industry salaries are significantly higher. Universities should explore means of providing internal support for data science staff, perhaps sharing capacity between academic groups or departments. Long-term vision and leadership in the field are needed to recognize and measure relevant metrics and make them part of advancement/career ladders.

8.5.2. *Science/Data centers*

At data centers, project data management (DM) teams need to cover a wide range of expertise such as astronomical domain knowledge, strong astronomical data understanding, deep software engineering skills and what is often referred to as “dev-ops” skills (engineering, deploying and operating production services). Given the broad areas of competency required, a team with a couple of people (or worse, sub-teams) in each area of expertise quickly exceeds the "optimal team size" which means the team gets mired in communication overheads, has difficulty forming a common purpose and loses agility (including over-planning, inability to respond to shifting requirements or technologies, and makework to compensate for inhomogeneities in the division of labor). A hybrid team is one that is not only multi-disciplinary in constitution but consists of generalists with fluency in more that one domain.

By assembling hybrid teams that not only bring domain specialty but share a common understanding of other areas in the team’s competence sphere, it is possible to constrain a team to a manageable size; avoid over-division of labor and the fractioning of individuals work assignments, and reap the ability of multi-disciplinary teams to reach new, overarching insights into their problem space. Developing these hybrid teams includes supporting tech savvy researchers who have expertise in both the domains of astrophysics and software engineering or other data support skills.

Ultimately, supporting these hybrid teams requires investment in job stability. Longer-term grants aimed at building and supporting abiding, professional (non-student) data science capacity.

REC-25 Recognize software as part of the career path.

Area: Workforce. **Audience:** Manager . **Term:** Short

Software should be recognized in hiring and career development as a core product of modern astronomical research. Software outputs should be considered in all aspects of *academic performance appraisals, career applications, and promotion and tenure review cases.*

REC-26 Partnerships to support data science staff.
Area: Workforce. **Audience:** University . **Term:** Medium
Universities should explore means of providing support for data-science faculty and staff, perhaps sharing capacity between academic groups or departments internally or partnerships outside the university.

REC-27 Support long-term technical capacity.
Area: Workforce. **Audience:** Agency . **Term:** Medium
Funding agencies should explore longer-term grants aimed at building and supporting professional (non-student) data science capacity.

8.6. *Elevating the role of software as a product of the research enterprise*

Software is a critical part of modern research and yet there is generally poor support across the scholarly ecosystem for its acknowledgment and citation, and in turn, for measuring its impact. The majority of academic fields rely on a one-dimensional credit model whereby academic articles (and their associated citations) are the dominant factor in the success of a researcher's career.

In the petabyte era of astronomical science, making it possible to easily cite software and measure its impact is going to be critical for maximizing the scientific return of these large datasets and retaining those individuals who specialize in developing the tools to turn them into publications.

Evolving beyond the one-dimensional credit model requires overcoming several key challenges including the current scholarly ecosystem and scientific culture issues. Career paths for staff, including scientific staff, in these technical roles need to have clearly defined metrics and requirements that take into account how they are required to spend their time in support of the scientific enterprise.

The ecosystem around the publishing of scholarly work has not been set up to properly account for contributions to scientific discoveries made through tools, services and other infrastructure. Changes for the modern way in which science is done need to be made. Publications, like ApJ and AJ, are run by the AAS, a professional society, and are answerable to their boards that are elected by and comprise the membership of professional researchers, who also publish in them. Therefore, it is important to educate the larger community on changes that need to be made to support modern recognition standards for software services and then advocate for these changes with professional societies.

Social and cultural issues within the field also must be changed to normalize the appropriate acknowledgment of those who write software and support other science infrastructure tools. We need academics in positions of power (e.g. on promotion and tenure review committees, recruitment teams, grant review panels) to value software as an important product of research. Although change takes time, it is important that we begin making those changes with concrete and practical suggestions that can be incrementally introduced into accepted procedures and communal norms. These suggestions include the identification of metrics that support proper assessment of the impact of software on achieving scientific results. In recent years, substantial improvements to enable the citation of software and tracking of these citations has been made in astronomy and astrophysics.

8.6.1. *Measuring/citing the impact of software*

One key factor for improving the recognition of software within academia is to enable native software citation, that is, make it possible and required for authors to cite the software packages they have used in the process of carrying out their research, and to then count these citations in tools such as the Astrophysics Data System (ADS). Enabling software citation is both a technical challenge and a cultural one: recommendations for what software should be cited and when to cite it have been explored in community-wide effort at FORCE11 (Smith et al. 2016), and follow-on efforts are exploring some of the more technical aspects of how to implement these recommendations (11 (2019)).

Within astronomy and astrophysics, the Asclepias project[22] — a collaboration between AAS publishing, ADS, and the Zenodo data archive (hen 2017) — is working to enable first-class support for software citation in AAS journals as well as support for indexing (counting) these citations within ADS. While this project is currently scoped to AAS journals only, the changes being made to support the citation and indexing of software serve as an example for other journals to follow suit.

8.6.2. *Strategies for elevating the role of software*

Part of the challenge of elevating the role of software within academia is to find actionable changes that improve the career prospects of those individuals writing research software. In this section, we outline a number of possible approaches.

Software papers: One approach gaining traction across a number of research disciplines is to allow papers about software to be published in "conventional" journals alongside other research papers, thereby making software more visible to the academic community, and giving software engineers a citable "creditable" entity (a paper) to include on their resume. Examples of journals within astronomy that demonstrate a willingness to follow this approach include PASP[23] and AAS publishing, which recently changed its editorial policies to explicitly allow software papers in their publications (AAS Publishing 2015). More recently AAS publishing has announced a partnership with another journal specializing in software review (Vishniac & Lintott 2018).

Enabling support for software citation and indexing: Another key factor in raising the visibility of research software is to enable software citation, count these citations, and then make these metrics visible to the world. As part of the work of the Asclepias project, software citations are not only being counted in the astronomical literature, they are also being made visible on the ADS website next to the paper record on ADS.

[22] http://adsabs.github.io/blog/asclepias
[23] https://iopscience.iop.org/journal/1538-3873

Inform and educate the community about software contributions: Organizations play a critical role in improving the career prospects of those writing research software as they are responsible for hiring these individuals, evaluating their performance, and making decisions about possible promotions/career advancement. One immediately actionable approach is to encourage prospective employees and current staff to list software they have developed on their resumes and performance appraisals. This would allow review committees to include software as part of their evaluations.

Community prizes: AAS has a collection of prizes for scientific merit, instrumentation, education, and service to the field. As it is an important part of scientific discovery, software contributions that have had a lasting positive impact on the field, should also be recognized with a new dedicated prize and/or as a recognized example of merit within these other prize categories.

Grants: The amount of research funding secured is an established metric for evaluating an individual. As recommended in Section 6, allowing existing funding streams to be utilized for software development provides a simple mechanism for funding research software, but also signaling community recognition for the impact and relevance of the individual writing this software. Furthermore, widespread availability of grant funding in support of software development would provide a strong incentive for universities to hire technical astronomers into tenure track positions.

REC-28 Adopt best practices for software citation.
Area: Workforce. **Audience:** Astronomer . **Term:** Short
Journals and reviewers should adopt best practices for assuring that software and other science support infrastructure is properly referenced and cited in articles. Referees and other reviewers should be trained to recognize when such acknowledgement is necessary and ask authors to provide that information.

REC-29 Adopt promotion metrics that acknowledge software and other science support.
Area: Workforce. **Audience:** Manager . **Term:** Long
Departments and other members of the community should adopt and use suggested metrics for promotion and tenure reviews of those scientists whose work and contributions involve software and science infrastructure.

REC-30 Community prizes for software contributions.
Area: Workforce. **Audience:** Agency . **Term:** Short
Professional astronomy societies should create dedicated prizes and allow for software contributions to be recognized as a criteria of merit within existing prizes.

9. A NEED FOR DEDICATED EDUCATION AND OUTREACH EXPERTISE

Contributors: *Amanda E. Bauer <abauer@lsst.org>, Britt Lundgren, Meg Schwamb, Brian Nord, Dara J Norman*

Note: If you have come directly to this chapter we suggest you please read at least the Introduction in Section 1 before delving further.

We need to capitalize on positive trends in digital literacy, the increasing use of mobile devices, and a discovery space driven by social media, through the progressive development of online resources in astronomy education and public outreach (EPO). The goal for this chapter is to clarify and bolster the multitude of opportunities that exist to develop newly accessible online tools to engage fellow citizens in the era of petabyte-scale astronomy.

Maintaining support for astronomy research relies on our ability to effectively communicate our science and cultivate public excitement and engagement. Historically, strategic programming for astronomy EPO in science projects has been an afterthought: the work has primarily been undertaken by astronomers who are passionate about EPO but may lack the specific professional skills required to do it effectively at scale. Moreover, most astronomers are not compensated for their time or rewarded by their efforts in EPO. To maximize the public impact of large projects in the petabyte era, we must give professional credit to astronomers who do outreach work and also dedicate resources to full-time personnel to develop, execute, and evaluate modern EPO activities.

Traditional means of public engagement (e.g. classroom visits, online videos, public lectures and panels, etc.) have demonstrated their importance and value, and have carved a niche in the landscape of public engagement. However, we have entered a new era of technology and social interaction, which necessitates new modalities for innovative pedagogical techniques, communication, and even scientific exploration. Taking advantage of opportunities of modern technology requires putting in place the appropriate professionals to create and develop the interfaces and connections to curricula that maximize adaptability and use. For example, connecting non-experts with ever larger datasets requires educators who have astronomy domain expertise (to curate and work with datasets) as well as expertise in innovative pedagogical practices.

In this new era of engagement, EPO teams who develop ground-breaking activities and pedagogical frameworks will have started the design process as early as possible (including during construction of new facilities) and will have drawn on a number of areas of expertise: astronomical research methods, educational theory and practice, web development and design, software engineering, and multi-modal communication.

In this chapter, we discuss recommendations and effective practices for advancing astronomy in society through data-driven education and outreach activities for maximizing the impact large observing facilities and data centers will provide. We begin by discussing the creation of accessible online activities (Section 9.1), then identify a range of skills needed to create such activities (Section 9.2), and finally, we establish the benefits of resourcing dedicated EPO groups from the earliest stages of astronomy facility planning and including EPO as part of the mission of projects (Section 9.3).

9.1. *Create accessible online activities for the public*

REC-31 Create accessible online Activities for the Public.
Area: EPO. **Audience:** Educator, Astronomer. **Term:** Short
To maximize the impact of astronomy in society in the rapidly approaching petabyte and exabyte eras, we recommend that projects and data centers develop accessible web interfaces and tools that enable the public to access, explore, and analyze authentic astronomical data at unprecedented scale.

Many good arguments have been made for enabling non-professionals and students to access and engage with authentic data and professional tools. However, in practice, the increasing complexity of interfaces to large datasets can become a barrier to access and use. User interfaces need to be attractive and intuitive for non-specialists and usable from mobile devices and platforms commonly used in schools (such as chromebooks and tablets). Interfaces created for professionals do not necessarily work for non-specialists, because they tend to have the following characteristics: 1) offer too many options; 2) do not offer a clear path toward a learning outcome; 3) too slow, unresponsive, or burdensome for the internet connections. Effort should be spent on user interfaces for public audiences, and ideally, on creating introductory activities as preparation for more complicated tasks.

Surveying users to assess their needs and interests helps the content design process and continues to improve the quality of an experience for users when a program in running. User testing is a regular practice for many companies that deliver a product to the public and is a process that should be adapted within astronomy EPO programs to ensure activities remain relevant and useable.

9.1.1. *Examples of online activities*

Several examples of existing and planned infrastructures illuminate avenues for online public engagement: below, we discuss Sloan Digital Sky Survey's (SDSS) SkyServer, Zooniverse's Citizen Science, NASA's Universe of Learning, and the EPO program of LSST.

For over 15 years, the SDSS has made its vast database of imaging and spectroscopic data (~200 TB) freely available to the world. The web-based SDSS data browser, SkyServer[24], provides a public entry point for navigating the data online. The numerous and diverse query and analysis tools available through the SkyServer are designed to meet the needs of astronomers and non-professionals alike. The benefit to this design is that any interested student or member of the public has unrestricted access to research-grade inquiries and applications of the data. However, the large number of available features and the technical jargon that accompany them often overwhelm non-experts, as well as professional astronomers who are external to the SDSS collaboration.

[24] http://skyserver.sdss.org

In order to better support audiences who may be put off or overwhelmed by the professional-grade access points to the SDSS database, the SDSS Education and Public Outreach team developed activities with simplified query tools and smaller, curated datasets to facilitate activities for pre-college educators and students (e.g. SDSS *Voyages*[25]). For non-specialist audiences, these activities lower the barrier to accessing the same authentic data, while providing an introduction to concepts related to both astronomy and data structures. For students and educators who may be interested in using the data for more advanced explorations, SDSS Voyages provides a helpful stepping stone. The next two sections of this chapter suggest avenues to promote this transition in other ongoing and planned astronomy projects and facilities.

Citizen science represents an example of successful use of the modern age of web connectivity by directly engaging the public in scientific research. Online citizen science enables scientists to work with the general public to perform data-sorting and analysis tasks that are difficult or impossible to automate, or that would be insurmountable for a single person or for small groups of individuals to undertake (Marshall et al. 2015). Highly accessible citizen science activities can advance both science and learning in the era of large astronomical datasets. Moreover, most participants from the public claim that the main reason they participate is the contribution they are making to fundamental science research (Cox 2017). Through online citizen science portals such as the Zooniverse[26] (Lintott et al. 2011), millions of volunteers have participated directly in this collaborative research experience, contributing to over 70 astronomy-based research papers. Another reason for the continued success of the Zooniverse platform in particular, is that it looks good and feels modern, even after a decade of activity. While professional astronomers are the PI's of citizen science projects, Zooniverse employs 13 developers, one designer, and two postdocs to lead the infrastructure development of the platform between the Adler and Oxford locations.

Members of the Zooniverse team have furthered the project's educational impact by developing a college-level data science curriculum around their crowd-sourced data. The NSF-funded Improving Undergraduate STEM Education (Improving Undergraduate STEM Education (IUSE)) Project: "Engaging Introductory Astronomy Students in Authentic Research through Citizen Science" (PI: L. Trouille) is a particularly successful example of scoping big-data astronomy for a college non-major audience. This innovative curriculum equips students with the essential tools to explore the intrinsic and environmental properties of 20,000 low-redshift SDSS galaxies that have morphological classifications from Galaxy Zoo. This project utilizes a curated dataset in Google Sheets and a simple, plug-in tool that enables intuitive data cropping and visualization. Instead of learning about galaxies and cosmology through traditional readings and lectures, students are challenged to discover key patterns and properties of the universe themselves, through first-hand explorations of authentic astronomical data. In the process, they gain skills in quantitative analysis, improve their overall data literacy, and practice science communication. The curriculum specifically

[25] http://voyages.sdss.org
[26] http://www.zooniverse.org

provides an opportunity to discuss the complications and limitations of authentic data, and the challenges of framing a question that can be effectively tested with the data one has in hand. The project is a great case study of delivering specific, high-impact learning outcomes through an analysis of authentic data, without requiring students to navigate full-scale datasets or jargon-rich professional tools for visualization and analysis.

NASA's Universe of Learning[27] offers a variety of individual online web pages that are well presented. A potential challenge for a typical user who finds one of these pages is knowing what to do next. Beyond exploring the beautiful multi-wavelength images space telescopes provide, there is not a clear path for a user to navigate toward specific learning outcomes or experiences.

LSST's EPO program[28] is unique among ground-based telescope projects: not only is it being constructed in tandem with the physical observatory itself, but the outreach program is funded at 2% of the project cost. EPO products will go live when the LSST Survey begins in 2022. EPO products were included from the beginning as part of the construction Project deliverables, because they faced similarly unique challenges as the data resulting from the survey itself. During its design phase, the EPO team selected specific audiences and invested in user needs assessments to examine what these audiences want, and cannot find elsewhere. Some major findings include the necessity for mobile-friendly interfaces, a clear path toward learning goals, and educators needing no new software to download in order to introduce classroom activities. This has shaped the overall strategy for LSST EPO development and the skill sets needed on the EPO Team, which is a small, interdisciplinary team of astronomers, writers, designers, educators, and developers. The mission of LSST EPO is "to offer accessible and engaging online experiences that provide non-specialists access to, and context for, LSST data so anyone can explore the Universe and be part of the discovery process."

The operations website will feature news about LSST discoveries, profiles of LSST scientists and engineers and their work, and will be optimized for use on mobile devices. The EPO team is also developing online, data-driven classroom investigation activities for students in advanced middle school through college. The topics cover commonly-taught principles in astronomy and physics, and each investigation is designed for use with Next Generation Science Standards (NGSS) in the United States and the Curriculum Nacional in Chile. All investigations come with support and assessment materials for instructors and no special software is needed to access the investigations, which will be available in English and Spanish. LSST EPO will maintain an easy-to-use gallery of high-quality multimedia visualizations that can be downloaded and integrated into exhibits and presentations. Finally, LSST EPO will provide support to researchers who create Citizen Science projects using LSST data, including a dedicated project-building tool on the Zooniverse platform. The infrastructure to host these activities is being built during construction and will take several years. Another critical task during construction is building prototypes and performing

[27] https://www.universe-of-learning.org/
[28] https://www.lsst.org/about/epo

user testing, which has continually proven to improve the user experience and usability of interfaces.

A consistent theme that emerges when examining these examples is that well-defined learning outcomes for activities, curated access to authentic data, and simple, intuitive design are important to prepare for the EPO response to the large data we will collect in the 2020s. The remaining sections identify areas of expertise EPO teams can employ to achieve these outcomes.

9.2. *Bring expertise to astronomy education and outreach teams*

REC-32 Bring dedicated experts onto astronomy education and outreach teams.
Area: EPO. **Audience:** Manager, Educator. **Term:** Medium
To create the accessible online interfaces that maximize public impact in the next decade, we recommend supporting dedicated education and outreach teams that pair astronomers with technical and education specialists to increase relevance, adoptability, and accessibility of activities.

The large-scale data challenges that face astronomy described in this paper also represent challenges and opportunities for formal education, public outreach, and science communication. A natural instinct for astronomers may be to adapt their new computational experience to outreach efforts. This is a noble goal, but astronomers are not be expected to know effective practices around mobile-friendly development, intuitive user interfaces for the public, marketing through social media, or how to connect astronomy activities to formal education curriculum standards. A team of EPO experts can advise and assist with these areas, which are essential to build successful activities that are discoverable and adoptable.

We recommend astronomy organizations support creating EPO teams with expertise in relevant areas. It is understood that to reach maximal impact of outreach activities, these individuals work with astronomers to combine astronomy and data science expertise with specific EPO expertise and experience. This section describes options for areas of expertise and roles that can be brought on to achieve specific goals.

Educators in the United States (US) are currently required to submit paperwork to demonstrate that they are teaching specific topics related to curriculum standards. An EPO **education specialist or instructional designer** brings knowledge of relevant curriculum standards and rubrics (for example, the Next Generation Science Standards[29]) and is able to connect astronomy activities to topics educators must cover. This is the most relevant for K-12 formal education in a traditional setting or homeschooling. An educational specialist can build professional development programs to increase confidence for bringing such activities into their classrooms if there is not an expert available to join in person.

An education specialist can also tap into educator networks to advertise existing programs and perform professional development. An example is the National Science Teachers Association (National Science Teachers Association (NSTA)) annual meeting and AAS.

[29] NGSS: https://www.nextgenscience.org/

An education specialist working with an astronomer can create curate datasets to achieve specific learning outcomes without overwhelming non-specialists (Rebull et al. 2018).

An **Information Officer or Communications Manager** can act as a primary contact for a facility and also set overall communication strategies and implementation plans. Topics covered in such strategies could include audiences, content priorities, communication channels, messaging, procedures, and more.

An **Outreach Specialist** could serve a range of purposes depending on the needs of the group. This could be a science writer, someone who responds to and directs questions received from audience groups, or contributes to social media presence. or an astronomer trained in science communication. If this person has astronomy training, he/she could work with astronomical datasets to curate options for the public.

Social Media is becoming increasingly important as a source of news and information in society. Dedicating a full-time equivalent (or more) to the role of **social media engagement specialist** increases awareness of EPO activities and engages various audiences to participate with activities that exist.

Overall branding, the look and feel of online activities, and developing interesting graphics and images to support press releases or other activities are the role of a **Graphic Designer**.

An **evaluation specialist** informs methods for understanding the impact of programs on specific audience groups. The most benefit occurs when the method for evaluating the success of a program is built into the development of the program itself. Metrics could include and are not limited to web analytics, short or long surveys, interviews, login requests, focus groups, web submission forms, and social media interactions.

A **web developer** considers the user interface and experience when visiting a site. Mobile-friendly accessibility is a requirement for non-specialists since most users of an online interface will discover the materials via social media and will access them from a mobile device, not a desktop platform. In addition, the most common machines used by schools are chromebooks and potentially weak internet connections, which require lightweight design. Development needs to satisfy these requirements are best implemented by experts in the field.

A **Software Architect** designs, deploys, and maintains production services for an online program. It is important to not overburden internet systems that can be common in classroom settings or non-urban areas.

A **Project Manager** oversees the detailed budget, schedule, contracts, documentation, and reporting. This role is important for programs being built during the construction of an astronomical facility.

9.3. *Fund dedicated astronomy education and outreach groups*

REC-33 Fund dedicated or centralized astronomy education and outreach groups.
Area: EPO. **Audience:** Agency. **Term:** Long
We recommend that funding agencies supporting the development and operation of large astronomical observing and data facilities fund professional education and outreach groups

who can provide strategy, oversight, and best practices to maximize the impact of outreach efforts, and encourage EPO efforts to be part of a project's mission.

Having a dedicated individual or team to develop the EPO program for a specific facility can improve efficiency, impact, and cost effectiveness. Strategic planning provides an opportunity to emphasize the identity of a particular large facility; to identify non-specialist audiences who could benefit the most from dedicated engagement; put into place best practices in outreach and communication programs; and complement the overall landscape of astronomy EPO efforts. It is important that the EPO professionals are employed directly at professional telescope facilities in order to emphasize the uniqueness of the program, build and maintain relationships with those doing the technical and scientific work, and help handle the astronomy-specific data products that currently require a reasonable level of understanding to interpret and use (see Section 4).

A dedicated EPO team also serves as a resource for enabling astronomers working with large datasets and data facilities to do more impactful and wide-reaching outreach. Groups that are specifically charged to do EPO can improve the impact of the existing NSF Broader Impacts investment by supporting astronomers to tap into existing programs. This improves discoverability of the EPO work astronomers are doing increases the likelihood of achieving Broader Impact goals at both the individual and NSF levels.

An EPO team could provide any of the following benefits:

- Conducting science communication and media training sessions for astronomers doing these activities.
- Providing introductions to various social media platforms that can be used for unique outreach experiences.
- Marketing and promoting activities through established social media and common online training resources (e.g. Code Academy).
- Creating or tapping into a centralized repository for people looking for resources
- Creating opportunities for collaboration between astronomers and existing outreach infrastructure that will promote success and provide wide-reaching impact. Examples include *Journey Through the Universe* in Hawai'i or *AstroDay* in Chile, both led by Gemini EPO.
- Performing user needs assessments and user testing to improve the quality of existing activities and to develop new programs that meet the needs of specific audiences.
- Evaluating and reporting on the impact of EPO activities. Evaluation methods can and should be built into program design.
- Providing guidance for astronomers when developing science drivers and use cases for educational materials and public interfaces related to their research expertise.
- Broadening participation to non-traditional audience groups.

The timing for building expert EPO teams should occur during the construction of a new facility and be included as part of the project's mission. Starting early affords time to implement appropriate strategy and infrastructure. Educational materials, supplemental professional development materials, striking visualizations and images and communications strategies should be ready at the start of a project to maximize the public impact of the facility.

In this chapter, we discussed recommendations and effective practices that can be employed to maximize the impact of large astronomy facilities and data centers in the next decade. We prefaced the need for creating accessible online astronomy activities for the public and identified a range of skills needed to create such activities. Finally, we established the benefits of resourcing dedicated EPO groups from the earliest stages of astronomy facility planning and even including EPO as part of the mission of projects.

These recommendations are based on two main things: trends seen elsewhere on the web that successfully respond to this new era of technology and social interactions on the web, and case studies within astronomy that demonstrate appropriate avenues for increasing engagement and accessibility through online activities.

Note: This chapter is the basis for an Astro2020 Decadal Survey APC white paper which will go into detail on resourcing and prioritizing recommendations. If you are interested in commenting or constributing, please contact Amanda E. Bauer <abauer@lsst.org>.

10. CONCLUSION

We had a good opportunity to think about a range of topics which have been detailed in this document. Several APC white papers will be published using this as a basis. If you are interested in contributing to or endorsing https://tinyurl.com/y2ksemp2[30] or contact the authors listed in this document. We do not intend this to be the solution to all issues rather a discussion of potential ways forward for the next decade. We shall host a third and final workshop[31] in October 2019 which will explore practical approaches to dealing with some of the recommendations raised here.

ACKNOWLEDGMENTS

The workshops in this series were supported by The Kavli Foundation[32]. The Kavli Foundation is dedicated to advancing science for the benefit of humanity, promoting public understanding of scientific research, and supporting scientists and their work.

[30] https://tinyurl.com/y2ksemp2
[31] https://petabytestoscience.github.io/
[32] https://www.kavlifoundation.org/

Glossary

1D: One-dimensional. 42

2D: Two-dimensional. 42, 73

3D: Three-dimensional. 42

AAG: Astronomy and Astrophysics Research Grants, an NSF program. 30

AAS: American Astronomical Society. 33, 37, 60, 61, 62, 67

ADS: Astrophysics Data System. 33, 61

AI: Artificial Intelligence, used to cover many machine learning algorithms. 46, 74

AJ: The Astronomical Journal. 60

APC: activities, projects, or state of the profession considerations - wrt. the decadal survey. 5, 70, 71

API: Application Programming Interface. Usually this is either "web API" (meaning how you access/update data via http) or "software API" (meaning how you call the function/class/etc in a particular programming language or library), although often only "API" is used and the prefix is implicit.. 17, 21, 23, 26, 28, 74

APL: Apache Public License. 28

ATLAS: The Asteroid Terrestrial-impact Last Alert System. 11

CADC: Canadian Astronomy Data Centre. 16

CAOM: Common Archive Observation Model http://www.opencadc.org/caom2/. 16

CI: cyberinfrastructure. 23, 24, 25, 72

Citizen Science: - Amanda, Arfon, Meg?,. 64, 66

cloud: A visible mass of condensed water vapor floating in the atmosphere, typically high above the ground or in interstellar space acting as the birthplace for stars. Also a way of computing (on other peoples computers leveraging their services and availability).. 16, 17, 18, 19, 20, 21, 23, 25, 27, 48, 49

CMB: Cosmic Microwave Background. 13

cold storage: Data moved to cold storage means data moved to cheaper slower storage such as tape. The assumption is this is no longer accessed frequently.. 19, 72

community software: Software developed for and shared among a large group of relatively like-minded users (e.g. astronomers). Typically, but not necessarily, open source software and open development-based.. 30, 31, 35, 36, 38, 39

container: A container is a software package that contains everything the software needs to run. This includes the executable program as well as system tools, libraries, and settings. ... For example, a container that includes PHP and MySQL can run identically on both a Linux computer and a Windows machine.. 28, 72, 73

CSSI: Cyberinfrastructure for Sustained Scientific Innovation https://www.nsf.gov/pubs/2019/nsf19548/nsf19548.htm. 36

cyberinfrastructure: Sometimes denoted CI, A term first used by the US National Science Foundation (NSF), and it typically is used to refer to information technology systems that provide particularly powerful and advanced capabilities.. 21, 22, 23, 24, 25, 28, 72

DES: Dark Energy Survey. 9

DESI: Dark Energy Spectroscopic Instrument. 9, 10, 11, 32, 33

DIBBs: Data Infrastructure Building Blocks. 36

DKIST: Daniel K. Inouye Solar Telescope (formerly the Advanced Technology Solar Telescope, ATST). 32, 33, 39

DM: Data Management. 27, 59

Docker: A popular implementation of container technology.. 23, 26, 28, 32, 34

DOE: Department Of Energy. 17, 33

DOI: Digital Object Identifier https://www.doi.org/. 34

DVCS: Distributed Version Control System, a form of version control where the complete codebase - including its full history - is mirrored on every developer's computer. 36, 73

ELT: Extremely Large Telescope. 11

EPO: Education and Public Outreach. 5, 19, 63, 64, 66, 67, 68, 69, 70

ESA: European Space Agency. 74, 75

ESAC: European Space Astronomy Centre. 16

FITS: Flexible Image Transport System, is an open standard defining a digital file format useful for storage, transmission and processing of data. Files are formatted as tables or 2D images. 16

FORCE11: FORCE11 is a community of scholars, librarians, archivists, publishers and research funders interested in the Future of Research Communications and e-Scholarship. 61

FPGA: Field Programmable Gate Array, an integrated circuit which is fairly easily configurable.. 24

git: The most widely used DVCS software. 36

GPL: GNU Public License. 28

GW: Gravity Wave. 11

HPC: High Performance Computing. 2, 17, 24

HTC: High Throughput Computing. 2, 24

IAM: Identity and Access Management. 3, 25

IDL: Interactive Data Language, a programming language used for data analysis. Harris Geospatial [33]. 37, 38, 46

interoperability: the ability of systems or software to exchange and make use of information between them.. 10, 13, 21, 23, 26

[33] https://www.harrisgeospatial.com/Software-Technology/IDL

IPAC: No longer an acronym; science and data center at Caltech. 16

IRAF: Image Reduction and Analysis Facility, a collection of software written at the National Optical Astronomy Observatory (NOAO) geared towards the reduction of astronomical images in pixel array form.. 28, 37, 46

IRSA: Infrared Science Archive. 16

IUSE: Improving Undergraduate STEM Education. 65

IVOA: International Virtual-Observatory Alliance. 16, 17

JWST: James Webb Space Telescope (formerly known as NGST). 32, 33, 39

LIGO: The Laser Interferometer Gravitational-Wave Observatory. 11, 24

LISA: Laser Interferometer Space Antenna - European Space Agency (ESA) mission for 2030's. 11

LSST: Large Synoptic Survey Telescope. 9, 10, 11, 12, 13, 16, 24, 26, 27, 29, 31, 32, 33, 39, 44, 55, 64, 66

LSSTC: LSST Corporation, a not for profit organisation associated with LSST. 56

MAST: Mikulski Archive for Space Telescopes. 16

MCMC: Markov Chain Monte Carlo, a class of algorithms for sampling from a probability distribution.. 44

ML: Machine Learning (see also AI). 13, 42

MSE: Maunakea Spectroscopic Explorer. 10

NAS: Neural Architecture Search. 49

NASA: National Aeronautics and Space Administration. 16, 33, 64, 75

NASA ROSES: Research Opportunities in Earth and Space Science. 30

NGSS: Next Generation Science Standards https://www.nextgenscience.org/. 66

NOAO: National Optical Astronomy Observatories (USA). 57, 74

NSF: National Science Foundation. 17, 30, 33, 36, 69, 72

NSTA: National Science Teachers Association. 67

NYU: New York University. 51

ODBC: Open DataBase Connectivity, a standard API for SQL databases.. 23

open development: A process for developing software that emphasizes all code contribution and decision-making be done in the open, available to as wide a group as possible (This usually means anyone with internet access).. 35, 36, 72, 74

open source software: Open source software is a type of software in which source code is released under a license in which the copyright holder grants users the rights to study, change, and distribute the software to anyone and for any purpose. Note that this is *not* necessarily the same as open to contribution (see open development).. 32, 33, 35, 72

OpenEXR: a high dynamic range raster file format, released as an open standard along with a set of software tools created by Industrial Light & Magic (ILM) http://www.openexr.com/index.html. 16

OS: Operating System. 27

Pan-STARRS: Panoramic Survey Telescope and Rapid Response System. 9

parquet: Parquet File Format Hadoop. Parquet, an open source file format for Hadoop. Parquet stores nested data structures in a flat columnar format. Compared to a traditional approach where data is stored in row-oriented approach, parquet is more efficient in terms of storage and performance.. 16, 26, 75

PASP: Publications of the Astronomical Society of the Pacific. 33

PB: PetaByte. 26

PDF: Probability Density Function. 6

PHP: a popular general-purpose scripting language that is especially suited to web development.. 72

PI: Principle Investigator. 51, 65

PLATO: PLAnetary Transits and Oscillations of stars, the third medium-class mission in ESA's Cosmic Vision programme.. 10

PSF: Point Spread Function, describes the response of an imaging system to a point source or point object.. 46

reproducibility: (this one should have many definitions and we have to say WHICH version we are talking about) The ability to combine the same code and data and get the same result, or the ability to use the same code with different data to enforce a result, or there may be others. 32, 47

S3: Structured, imperative high level computer programming language, used as implementation language for the Virtual Machine Environment (Virtual Machine Environment (VME)) operating system. 17, 23, 27

SDSS: Sloan Digital Sky Survey. 9, 10, 24, 27, 42, 64, 65

SKA: Square Kilometer Array. 9, 13

software: The programs and other operating information used by a computer.. 3, 4, 5, 6, 7, 9, 10, 11, 13, 19, 21, 23, 24, 25, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 41, 46, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 66, 72, 73, 74

SQL: Structured Query Language, for interrogating relation databases. 23, 74

STEM: Science, Technology, Engineering and Math. 65, 74

TAP: Table Access Protocol. 17

TB: TeraByte. 25, 27, 64

TESS: Transiting Exoplanet Survey Satellite, a space telescope for NASA's Explorer program. 10

TPU: Tensor Processing Unit , a proprietary type of processor designed by Google in 2016 for use with neural networks and in machine learning projects. 24

US: United States. 67, 72

version control: The management of changes to documents, computer programs, and other collections of information. Changes are usually identified by a number or letter code. Each revision is associated with a timestamp and the person making the change. Revisions can be compared, restored, and merged. 54, 55, 73

VME: Virtual Machine Environment. 75

VO: Virtual Observatory. 17

WFIRST: Wide Field Infrared Survey Telescope. 9, 10, 12, 13, 16, 26

ZTF: Zwicky Transient Facility. 9, 11

REFERENCES

2017, Asclepias - Capturing Software Citations in Astronomy (Zenodo)

11, F. 2019

AAS Publishing. 2015

Andreessen, M. 2011, online

Astronomy & Computing. 2013, https://www.journals.elsevier.com/astronomy-and-computing

Astropy Collaboration. 2019

Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33

Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, AJ, 156, 123

Baron, D., & Poznanski, D. 2017, MNRAS, 465, 4530

Bechtol, K., et al. 2019, Astro2020 Science White Paper: Dark Matter Science in the Era of LSST

Behroozi, P., et al. 2019, Astro2020 Science White Paper: Empirically Constraining Galaxy Evolution

Bektesevic, D., Mehta, P., Juric, M., et al. 2019, in American Astronomical Society Meeting Abstracts, Vol. 233, American Astronomical Society Meeting Abstracts #233, 245.05

Capak, P., et al. 2019, Astro2020 Science White Paper: Synergizing Deep Field Programs Across Multiple Surveys

Chang, P., et al. 2019, Astro2020 Science White Paper: Cyberinfrastructure Requirements to Enhance Multi-messenger Astrophysics

Chanover, N., et al. 2019, Astro2020 Science White Paper: Triggered High-Priority Observations of Dynamic Solar System Phenomena

Chary, R., et al. 2019, Astro2020 Science White Paper: Cosmology in the 2020s Needs Precision and Accuracy: The Case for Euclid/LSST/WFIRST Joint Survey Processing

Collberg, C., & Proebsting, T. A. 2016, Commun. ACM, 59, 62

Cooray, A., et al. 2019, Astro2020 Science White Paper: Cosmic Dawn and Reionization: Astrophysics in the Final Frontier

Cowperthwaite, P., et al. 2019, Astro2020 Science White Paper: Joint Gravitational Wave and Electromagnetic Astronomy with LIGO and LSST in the 2020's

Cox, P. 2017, in American Astronomical Society Meeting Abstracts, Vol. 229, American Astronomical Society Meeting Abstracts #229, 411.07

Cuby, J., et al. 2019, Astro2020 Science White Paper: Unveiling Cosmic Dawn: the synergetic role of space and ground-based telescopes

Dey, A., et al. 2019, Astro2020 Science White Paper: Mass Spectroscopy of the Milky Way

Dickinson, M., et al. 2019, Astro2020 Science White Paper: Observing Galaxy Evolution in the Context of Large-Scale Structure

Donoho, D. L., Maleki, A., Rahman, I. U., Shahram, M., & Stodden, V. 2009, Computing in Science & Engineering, 11, 8

Dore, O., et al. 2019, Astro2020 Science White Paper: WFIRST: The Essential Cosmology Space Observatory for the Coming Decade

Dvorkin, C., et al. 2019, Astro2020 Science White Paper: Neutrino Mass from Cosmology: Probing Physics Beyond the Standard Model

Eifler, T., et al. 2019, Astro2020 Science White Paper: Partnering space and ground observatories - Synergies in cosmology from LSST and WFIRST

Eracleous, M., et al. 2019, Astro2020 Science White Paper: An Arena for Multi-Messenger Astrophysics: Inspiral and Tidal Disruption of White Dwarfs by Massive Black Holes

Fabbiano, G., et al. 2019a, Astro2020 Science White Paper: Increasing the Discovery Space in Astrophysics The Exploration Question for Compact Objects

—. 2019b, Astro2020 Science White Paper: Increasing the Discovery Space in Astrophysics The Exploration Question for Cosmology

—. 2019c, Astro2020 Science White Paper: Increasing the Discovery Space in Astrophysics The Exploration Question for Galaxy Evolution

—. 2019d, Astro2020 Science White Paper: Increasing the Discovery Space in Astrophysics The Exploration Question for Planetary Systems

—. 2019e, Astro2020 Science White Paper: Increasing the Discovery Space in Astrophysics: The Exploration Question for Resolved Stellar Populations

—. 2019f, Astro2020 Science White Paper: Increasing the Discovery Space in Astrophysics The Exploration Question for Stars and Stellar Evolution

Fan, X., et al. 2019, Astro2020 Science White Paper: The First Luminous Quasars and Their Host Galaxies

Ferraro, S., et al. 2019, Astro2020 Science White Paper: Inflation and Dark Energy from spectroscopy at $z > 2$

Ford, E., et al. 2019, Astro2020 Science White Paper: Advanced Statistical Modeling of Ground-Based RV Surveys as Critical Support for Future NASA Earth-Finding Missions

Furlanetto, S., et al. 2019, Astro2020 Science White Paper: Synergies Between Galaxy Surveys and Reionization Measurements

Gaudi, B., et al. 2019, Astro2020 Science White Paper: "Auxiliary" Science with the WFIRST Microlensing Survey

Geach, J., et al. 2019, Astro2020 Science White Paper: The case for a 'sub-millimeter SDSS': a 3D map of galaxy evolution to z 10

Geiger, R. S., Cabasse, C., Cullens, C. Y., et al. 2018, Career Paths and Prospects in Academic Data Science: Report of the Moore-Sloan Data Science Environments Survey

GitHub. 2016, online

Gluscevic, V., et al. 2019, Astro2020 Science White Paper: Cosmological Probes of Dark Matter Interactions: The Next Decade

Gnau, S. 2017, online

Graham, M., et al. 2019, Astro2020 Science White Paper: Discovery Frontiers of Explosive Transients: An ELT and LSST Perspective

Grin, D., et al. 2019, Astro2020 Science White Paper: Gravitational probes of ultra-light axions

Hannay, J. E., MacLeod, C., Singer, J., et al. 2009, in 2009 ICSE Workshop on Software Engineering for Computational Science and Engineering (IEEE)

Hanselman, S. 2006, Sandcastle - Microsoft CTP of a Help CHM file generator on the tails of the death of NDoc

Holler, B., et al. 2019, Astro2020 Science White Paper: "It's full of asteroids!": Solar system science with a large field of view

Hornik, K. 2018, R FAQ

Huber, D., et al. 2019, Astro2020 Science White Paper: Stellar Physics and Galactic Archeology using Asteroseismology in the 2020's

Hunter, J. D. 2007, Computing In Science & Engineering, 9, 90

Ivezić, Ž., Connolly, A., Vanderplas, J., & Gray, A. 2014, Statistics, Data Mining and Machine Learning in Astronomy (Princeton University Press)

Jao, W.-C., Henry, T. J., Gies, D. R., & Hambly, N. C. 2018, The Astrophysical Journal, 861, L11

Jones, E., Oliphant, T., Peterson, P., et al. 2001–, SciPy: Open source scientific tools for Python, [Online; accessed 2019-03-14]

Katz, D. S., Druskat, S., Haines, R., Jay, C., & Struck, A. 2018, arXiv e-prints, arXiv:1807.07387

Kirkpatrick, J., et al. 2019, Astro2020 Science White Paper: Opportunities in Time-domain Stellar Astrophysics with the NASA Near-Earth Object Camera (NEOCam)

Kluyver, T., Ragan-Kelley, B., Pérez, F., et al. 2016, in Positioning and Power in Academic Publishing: Players, Agents and Agendas, ed. F. Loizides & B. Schmidt, IOS Press, 87

Kohl, M. 2015, Introduction to statistical data analysis with R (London: bookboon.com)

Kollmeier, J., et al. 2019, Astro2020 Science White Paper: Precision Stellar Astrophysics and Galactic Archeology: 2020

Kupfer, T., et al. 2019, Astro2020 Science White Paper: A Summary of Multimessenger Science with Galactic Binaries

LeClair, H. 2016, online

Lehner, N., et al. 2019, Astro2020 Science White Paper: Following the Metals in the Intergalactic and Circumgalactic Medium over Cosmic Time

Li, T., et al. 2019, Astro2020 Science White Paper: Dark Matter Physics with Wide Field Spectroscopic Surveys

Limoncelli, T. A. 2018, Queue, 16, 20:13

Lintott, C., Schawinski, K., Bamford, S., et al. 2011, MNRAS, 410, 166

Littenberg, T., et al. 2019, Astro2020 Science White Paper: GRAVITATIONAL WAVE SURVEY OF GALACTIC ULTRA COMPACT BINARIES

Maccarone, T., et al. 2019, Astro2020 Science White Paper: Populations of Black holes in Binaries

Mandelbaum, R., et al. 2019, Astro2020 Science White Paper: Wide-field Multi-object Spectroscopy to Enhance Dark Energy Science from LSST

Mantz, A., et al. 2019, Astro2020 Science White Paper: The Future Landscape of High-Redshift Galaxy Cluster Science

Marshall, P. J., Lintott, C. J., & Fletcher, L. N. 2015, ARA&A, 53, 247

McQueen, J., Meila, M., VanderPlas, J., & Zhang, Z. 2016, arXiv e-prints, arXiv:1603.02763

Meeburg, P., et al. 2019, Astro2020 Science White Paper: Primordial Non-Gaussianity

Momcheva, I., Smith, A. M., & Fox, M. 2019, in American Astronomical Society Meeting Abstracts, Vol. 233, American Astronomical Society Meeting Abstracts #233, 457.06

Momcheva, I., & Tollerud, E. 2015, arXiv e-prints, arXiv:1507.03989

Morris, K. 2016, Infrastructure as Code: Managing Servers in the Cloud, Safari Books Online (O'Reilly Media)

Moustakas, L., et al. 2019, Astro2020 Science White Paper: Quasar microlensing: Revolutionizing our understanding of quasar structure and dynamics

National Academies of Sciences, Engineering, and Medicine. 2018, Open Source Software Policy Options for NASA Earth and Space Sciences (Washington, DC: The National Academies Press)

Newman, J., et al. 2019, Astro2020 Science White Paper: Deep Multi-object Spectroscopy to Enhance Dark Energy Science from LSST

Ntampaka, M., et al. 2019, Astro2020 Science White Paper: The Role of Machine Learning in the Next Decade of Cosmology

Olsen, K., et al. 2019, Astro2020 Science White Paper: Science Platforms for Resolved Stellar Populations in the Next Decade

Palmese, A., et al. 2019, Astro2020 Science White Paper: Gravitational wave cosmology and astrophysics with large spectroscopic galaxy surveys

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825

Pevtsov, A., et al. 2019, Astro2020 Science White Paper: Historical astronomical data: urgent need for preservation, digitization enabling scientific exploration

Pisani, A., et al. 2019, Astro2020 Science White Paper: Cosmic voids: a novel probe to shed light on our Universe

Pooley, D., et al. 2019, Astro2020 Science White Paper: The Most Powerful Lenses in the Universe: Quasar Microlensing as a Probe of the Lensing Galaxy

Raymond, E. S. 2001, The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary (Sebastopol, CA, USA: O'Reilly & Associates, Inc.)

Rebull, L. M., French, D. A., Laurence, W., et al. 2018, Phys. Rev. Phys. Educ. Res., 14, 020102

Research Software Engineers International. 2018

Rhodes, J., et al. 2019a, Astro2020 Science White Paper: Cosmological Synergies Enabled by Joint Analysis of Multi-probe data from WFIRST, Euclid, and LSST

—. 2019b, Astro2020 Science White Paper: The End of Galaxy Surveys

Rix, H., et al. 2019, Astro2020 Science White Paper: Binaries Matter Everywhere: from Precision Calibrations to Re-Ionization and Gravitational Waves

Sanderson, R., et al. 2019, Astro2020 Science White Paper: The Multidimensional Milky Way

Shen, Y., et al. 2019, Astro2020 Science White Paper: Mapping the Inner Structure of Quasars with Time-Domain Spectroscopy

Slosar, A., et al. 2019a, Astro2020 Science White Paper: Dark Energy and Modified Gravity

—. 2019b, Astro2020 Science White Paper: Scratches from the Past: Inflationary Archaeology through Features in the Power Spectrum of Primordial Fluctuations

Smith, A. M., Katz, D. S., & and, K. E. N. 2016, PeerJ Computer Science, 2, e86

Smith, A. M., Niemeyer, K. E., Katz, D. S., et al. 2018, PeerJ Computer Science, 4, e147

SunPy Community, T., Mumford, S. J., Christe, S., et al. 2015, Computational Science and Discovery, 8, 014009

Suzuki, N., & Fukugita, M. 2018, AJ, 156, 219

van der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, Computing in Science & Engineering, 13, 22

Varoquaux, G. 2013, online

Vishniac, E. T., & Lintott, C. 2018, The Astrophysical Journal, 869, 156

von Hardenberg, A., Obeng, A., Pawlik, A., et al. 2019, Data Carpentry: R for data analysis and visualization of Ecological Data

Wang, Y., et al. 2019, Astro2020 Science White Paper: Illuminating the dark universe with a very high density galaxy redshift survey over a wide area

Weiner, B., Blanton, M. R., Coil, A. L., et al. 2009, in astro2010: The Astronomy and Astrophysics Decadal Survey, Vol. 2010, P61

Williams, B., et al. 2019, Astro2020 Science White Paper: Far Reaching Science with Resolved Stellar Populations in the 2020s

Wilson, G. 2013, arXiv e-prints, arXiv:1307.5448

Wilson, G., Aruliah, D. A., Brown, C. T., et al. 2014, PLOS Biology, 12, 1

Yale Roundtable Declaration. 2010, Computing in Science & Engineering, 12, 8

Yee, J., et al. 2019, Astro2020 Science White Paper: The Scientific Context of WFIRST, Microlensing in the 2020s