Draft version August 5, 2022
Typeset using LaTeX **modern** style in AASTeX62

From Data to Software to Science with the Rubin Observatory LSST

Katelyn Breivik,[1] Andrew J. Connolly,[2] K. E. Saavik Ford,[3,4,5,1] Mario Jurić,[2] Rachel Mandelbaum,[6] Adam A. Miller,[7] Dara Norman,[8] Knut Olsen,[8] William O'Mullane,[9] Adrian Price-Whelan,[1] Timothy Sacco,[8] J. L. Sokoloski,[10,11] Ashley Villar,[12,13,14] Viviana Acquaviva,[1,15] Tomas Ahumada,[16] Yusra AlSayyad,[17] Catarina S. Alves,[18] Igor Andreoni,[19,20,21] Timo Anguita,[22,23] Henry J. Best,[5] Federica B. Bianco,[24,25,26] Rosaria Bonito,[27] Andrew Bradshaw,[28,29] Colin J. Burke,[30] Andresa Rodrigues de Campos,[6] Matteo Cantiello,[1,17] Neven Caplar,[17] Colin Orion Chandler,[31] James Chan,[32] Luiz Nicolaci da Costa,[33] Shany Danieli,[17] James R. A. Davenport,[2] Giulio Fabbian,[1,34] Joshua Fagin,[5] Alexander Gagliano,[30] Christa Gall,[35] Nicolás Garavito Camargo,[1] Eric Gawiser,[36] Suvi Gezari,[37] Andreja Gomboc,[38] Alma X. Gonzalez-Morales,[39,40] Matthew J. Graham,[41] Julia Gschwend,[33] Leanne P. Guy,[9] Matthew J. Holman,[42] Henry H. Hsieh,[43] Markus Hundertmark,[44] Dragana Ilić,[45,46] Emille E. O. Ishida,[47] Tomislav Jurkić,[48] Arun Kannawadi,[17] Alekzander Kosakowski,[49] Andjelka B. Kovačević,[45] Jeremy Kubica,[6] François Lanusse,[50] Ilin Lazar,[51] W. Garrett Levine,[52] Xiaolong Li,[24] Jing Lu,[53] Gerardo Juan Manuel Luna,[54,55] Ashish A. Mahabal,[56,57] Alex I. Malz,[58,6] Yao-Yuan Mao,[36,59] Ilija Medan,[60] Joachim Moeyens,[2] Mladen Nicolić,[45] Robert Nikutta,[8] Matt O'Dowd,[61,4] Charlotte Olsen,[36] Sarah Pearson,[62] Ilhuiyolitzin Villicana Pedraza,[63] Mark Popinchalk,[64] Luka Č. Popović,[65,45] Tyler A. Pritchard,[62] Bruno C. Quint,[9] Viktor Radović,[45] Fabio Ragosta,[66] Gabriele Riccio,[67] Alexander H. Riley,[68,69] Agata Rożek,[70] Paula Sánchez-Sáez,[71] Luis M. Sarro,[72] Clare Saunders,[17] Đorđe V. Savić,[73,65] Samuel Schmidt,[74] Adam Scott,[8] Raphael Shirley,[75] Hayden R. Smotherman,[2] Steven Stetzler,[2] Kate Storey-Fisher,[62] Rachel A. Street,[76] David E. Trilling,[31] Yiannis Tsapras,[44] Sabina Ustamujic,[27] Sjoert van Velzen,[77] José Antonio Vázquez-Mata,[78,79] Laura Venuti,[80] Samuel Wyatt,[2] Weixiang Yu,[81] and Ann Zabludoff[82]

[1]*Center for Computational Astrophysics, Flatiron Institute, 162 Fifth Ave, New York, NY, 10010, USA*

[2]*Department of Astronomy and the DIRAC Institute, University of Washington, 3910 15th Avenue NE, Seattle, WA 98195, USA*

[3]*Department of Science, CUNY Borough of Manhattan Community College, 199 Chambers Street, New York, NY 10007, USA*

[4]*Department of Astrophysics, American Museum of Natural History, New York, NY 10028, USA*

[5]*The Graduate Center of the City University of New York, 365 Fifth Avenue, New York, NY 10016, USA*

[6]*McWilliams Center for Cosmology, Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15213, USA*

[7]*Center for Interdisciplinary Exploration and Research in Astrophysics (CIERA) and Department of Physics and Astronomy, Northwestern University, 1800 Sherman Road, Evanston, IL 60201, USA*

[8]*National Optical-Infrared Astronomy Research Laboratory, 950 N. Cherry Ave., Tucson, AZ 85719, USA*

[9]*Rubin Observatory Project Office, 950 N. Cherry Ave., Tucson, AZ 85719, USA*

[10]*The LSST Corporation*

[11]*Columbia University, 533 W. 218th St. New York, NY 10034*

[12]*Department of Astronomy and Astrophysics, The Pennsylvania State University, 525 Davey Lab, University Park, PA 16802, USA*

[13]*Institute for Computational & Data Sciences, The Pennsylvania State University, University Park, PA 16802, USA*

[14]*Institute for Gravitation and the Cosmos, The Pennsylvania State University, University Park, PA 16802, USA*

[15]*City University of New York, New York City College of Technology*

[16]*University of Maryland, College Park, MD 20742*

[17]*Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08544, USA*

[18]*University College London, Gower St, London WC1E 6BT, United Kingdom*

[19]*Joint Space-Science Institute, University of Maryland, College Park, MD 20742, USA*

[20]*Department of Astronomy, University of Maryland, College Park, MD 20742, USA*

[21]*Astrophysics Science Division, NASA Goddard Space Flight Center, Mail Code 661, Greenbelt, MD 20771, US*

[22]*Departamento de Ciencias Fisicas, Universidad Andres Bello Fernandez Concha 700, Las Condes, Santiago, Chile*

[23]*Millennium Institute of Astrophysics, Nuncio Monseñor Sótero Sanz 100, Of 104, Providencia, Santiago, Chile*

[24]*Department of Physics and Astronomy, University of Delaware, Newark, DE 19716-2570, USA*

[25]*Joseph R. Biden, Jr., School of Public Policy and Administration, University of Delaware, Newark, DE 19717 USA*

[26]*Data Science Institute, University of Delaware, Newark, DE 19717 USA*

[27]*INAF - Osservatorio Astronomico di Palermo, Piazza del Parlamento 1 90134, Palermo, Italy*

[28]*SLAC National Accelerator Laboratory, 2575 Sand Hill Rd., Menlo Park, CA 94025, USA*

[29]*Kavli Institute for Particle Astrophysics and Cosmology, SLAC*

[30]*University of Illinois, Department of Astronomy, 1110 W. Green St., Urbana, IL 61801, USA*

[31]*Department of Astronomy and Planetary Science, Northern Arizona University, P.O. Box 6010, Flagstaff, AZ 86011, USA*

[32]*City University of New York*

[33]*Laboratório Interinstitucional de e-Astronomia, Rua General José Cristino, 77, Rio de Janeiro, RJ, 20921-400, Brazil*

[34]*School of Physics and Astronomy, Cardiff University, The Parade, Cardiff, CF24 3AA, UK*

[35]*DARK, Niels Bohr Institute, University of Copenhagen, Denmark*

[36]*Department of Physics and Astronomy, Rutgers University, 136 Frelinghuysen Rd., Piscataway, NJ 08854, USA*

[37]*Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA*

[38]*Center for Astrophysics and Cosmology, University of Nova Gorica, Vipavska 13 5000 Nova Gorica, Slovenia*

[39]*Consejo Nacional de Ciencia y Tecnología, Av. Insurgentes Sur 1582. Colonia Crédito Constructor, Del. Benito Juárez C.P. 03940, México D.F. México*

[40]*Departamento de Física, DCI, Campus León, Universidad de Guanajuato, 37150, León, Guanajuato, México*

[41]*Astronomy Department, California Institute of Technology, 1200 East California Blvd., Pasadena CA 91125, USA*

[42]*Center for Astrophysics, Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138*

[43]*Planetary Science Institute, 1700 East Fort Lowell Road, Suite 106, Tucson, AZ 85719, USA*

[44]*Zentrum für Astronomie der Universität Heidelberg, Astronomisches Rechen-Institut, Mönchhofstr. 12-14, 69120 Heidelberg, Germany*

[45]*Faculty of Mathematics, Department of Astronomy, University of Belgrade, Studentski trg 16 Belgrade, Serbia*

[46]*Hamburger Sternwarte, Universitat Hamburg, Gojenbergsweg 112, 21029 Hamburg, Germany*

[47]*Université Clermont Auvergne, CNRS/IN2P3, Laboratoire de Physique de Clermont, F-63000 Clermont-Ferrand, France*

[48]*Faculty of Physics, University of Rijeka, Radmile Matejčić 2, Rijeka, Croatia*

[49]*Department of Physics and Astronomy, Texas Tech University, Lubbock, TX 79409, USA*

[50]*Université Paris-Saclay, Université Paris Cité, CEA, CNRS, AIM, 91191, Gif-sur-Yvette, France*

[51]*Centre for Astrophysics Research, University of Hertfordshire, Hatfield, Hertfordshire, AL10 9AB, UK*

[52]*Astronomy Department, Yale University, New Haven, CT 06520, USA*
[53]*Florida State University, Tallahassee, FL 32306*
[54]*Instituto de Astronomía y Física del Espacio (IAFE), Av. Inte. Güiraldes 2620, C1428ZAA, Buenos Aires, Argentina*
[55]*Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina*
[56]*Division of Physics, Mathematics and Astronomy, California Institute of Technology, Pasadena, CA 91125, USA*
[57]*Center for Data Driven Discovery, California Institute of Technology, Pasadena, CA 91125, USA*
[58]*German Centre for Cosmological Lensing, Astronomisches Institut, Ruhr-Universität Bochum, Universitätsstr. 150, 44801 Bochum, Germany*
[59]*Department of Physics and Astronomy, University of Utah, Salt Lake City, UT 84112, USA*
[60]*Georgia State University, Atlanta, GA 30302, USA*
[61]*Department of Physics and Astronomy, Lehman College, City University of New York, NY 10468, USA*
[62]*Center for Cosmology & Particle Physics, New York University, 726 Broadway, New York, 10003, USA*
[63]*DACC New Mexico State University, Central Campus, Las Cruces, New Mexico, USA*
[64]*City University of New York, American Museum of Natural History*
[65]*Astronomical Observatory, Volgina 7, P.O. Box 74, 11060 Belgrade, Serbia*
[66]*Istituto Nazionale di Astrofisica, Viale del Parco Mellini 84, 00136 Rome, Italy*
[67]*National Centre of Nuclear Research, Andrzeja Sołtana 7, 05-400 Otwock, Poland*
[68]*George P. and Cynthia Woods Mitchell Institute for Fundamental Physics and Astronomy, Texas A&M University, College Station, TX 77843, USA*
[69]*Department of Physics and Astronomy, Texas A&M University, College Station, TX 77843, USA*
[70]*Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK*
[71]*European Southern Observatory, Karl-Schwarzschild-Strasse 2, 85748 Garching bei München, Germany*
[72]*Dpt. of Artificial Intelligence, Universidad Nacional de Educación a Distancia, Madrid, Madrid, ES*
[73]*Institut d'Astrophysique et de Géophysique, Université de Liège, Allée du 6 Août 19c, 4000 Liège, Belgium*
[74]*Physics Department, University of California, One Shields Avenue, Davis, CA 95616, USA*
[75]*University of Southampton, Hartley Library B12, University Rd, Highfield, Southampton SO17 1BJ, United Kingdom*
[76]*Las Cumbres Observatory, 6740 Cortona Dr., Suite 102, Goleta, CA 93117, USA*
[77]*Leiden Observatory, Leiden University, Postbus 9513, 2300 RA, Leiden, The Netherlands*
[78]*Departamento de Física, Facultad de Ciencias, Universidad Nacional Autónoma de México, Ciudad Universitaria, CDMX, 04510, México*
[79]*Instituto de Astronomía sede Ensenada, Universidad Nacional Autónoma de México, Km 107, Carret. Tij.-Ens., Ensenada, 22060, BC, México*
[80]*SETI Institute, 339 Bernardo Avenue, Suite 200, Mountain View, CA 94043, USA*
[81]*Drexel University, 3141 Chestnut St, Philadelphia, PA 19104*
[82]*University of Arizona, Department of Astronomy and Steward Observatory, 933 N Cherry Ave, Tucson, AZ 85721, USA*

Abstract

The Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST) dataset will dramatically alter our understanding of the Universe, from the origins of the Solar System to the nature of dark matter and dark energy. Much of this research will depend on the existence of robust, tested, and scalable algorithms, software, and services. Identifying and developing such tools ahead of time has the potential to significantly accelerate the delivery of early science from LSST. Developing these collaboratively, and making them broadly available, can enable more inclusive and equitable collaboration on LSST science.

To facilitate such opportunities, a community workshop entitled "From Data to Software to Science with the Rubin Observatory LSST" was organized by the LSST Interdisciplinary Network for Collaboration and Computing (LINCC) and partners, and held at the Flatiron Institute in New York, March 28-30th 2022. The workshop included over 50 in-person attendees invited from over 300 applications. It identified seven key software areas of need: (i) scalable cross-matching and distributed joining of catalogs, (ii) robust photometric redshift determination, (iii) software for determination of selection functions, (iv) frameworks for scalable time-series analyses, (v) services for image access and reprocessing at scale, (vi) object image access (cutouts) and analysis at scale, and (vii) scalable job execution systems.

This white paper summarizes the discussions of this workshop. It considers the motivating science use cases, identified cross-cutting algorithms, software, and services, their high-level technical specifications, and the principles of inclusive collaborations needed to develop them. We provide it as a useful roadmap of needs, as well as to spur action and collaboration between groups and individuals looking to develop reusable software for early LSST science.

Author list: The organizing committee are listed in alphabetical order first. Other contributors and attendees are listed in alphabetical order after these. Endorsers, who did not contribute text but who support the overall goals and message of the white paper, are listed separately.
Editors: William O'Mullane and Mario Jurić

ENDORSERS

Camille Avestruz (University of Michigan), Massimo Brescia (University of Napoli Federico II), James J. Buchanan (Lawrence Livermore National Laboratory), Jeffrey L. Carlin (Rubin Observatory), Aleksandra Ćiprijanović (Fermi National Accelerator Laboratory), Kristen C. Dage (McGill University), Tansu Daylan (Princeton University), Mariano Dominguez (IATE-OAC-UNC and CONICET), Melissa L. Graham (University of Washington), Akhtar Mahmood (Bellarmine University), Martin Makler (ICAS & ICIFI, UNSAM - CONICET, Argentina and CBPF, Brazil), Eniko Regos (Konkoly Observatory), Bruno O. Sánchez (Duke University), Róbert Szabó (Konkoly Observatory), Christopher Theissen (UC San Diego), José Alberto Vázquez González (Instituto de Ciencias Físicas UNAM), Yuanyuan Zhang (Texas A&M University)

EXECUTIVE SUMMARY

The Vera C. Rubin Observatory and the Legacy Survey of Space and Time (formerly Large Synoptic Survey Telescope) (LSST) will dramatically alter our understanding of the Universe, from the origins of the Solar System to the nature of dark matter and dark energy. Many of the diverse science cases LSST will enable rely on the existence of robust, tested, and scalable algorithms and software. Concentrating on the development of a small set of crucial cross-cutting software components and services has the potential to enable many early science cases for the LSST science community.

This document presents the conclusions from a three day workshop entitled "From Data to Software to Science with the Rubin Observatory LSST", organized by LSST Interdisciplinary Network for Collaboration and Computing (LINCC) Frameworks working with Rubin Observatory and other partners such as Independent Data Access Centers (IDACs) and LSST Science Collaborations (SCs) with a goal of identifying key cross-cutting software components that can accelerate LSST science. The meeting was held at the Flatiron institute in New York from March 28th to 30th 2022. Due to logistical restrictions, attendance was restricted to around fifty in-person attendees, but plenaries were open through Zoom to the >300 applicants for the workshop. Care was taken to have diverse groups and all LSST SCs represented at the workshop, including participants from traditionally underrepresented groups and institutions. A focus of the meeting was the development of inclusive collaborations both within the LINCC initiative and more broadly in the SCs.

The meeting was structured around discussion of science use cases (Appendix B) that could be undertaken with the first 1-2 years of Rubin data. All use cases had a common structure, focusing on computational and software challenges that could limit the science community's ability to undertake their science. In total 41 use cases were developed across seven broad research areas: Solar System, Local Universe Static Science, Local Universe Variable and Transient Science, Extragalactic Static Science, Extragalactic Variable Science, Extragalactic Transient Science, and Cosmology.

From an evaluation of these science cases, six technical areas (Section 3) were identified where substantial development of cross-cutting software, algorithms, and computational infrastructure would need to be developed. These included: the ability to rapidly cross-match external data sets (e.g., at different wavelengths) with the Rubin alert stream and data release catalogs; the development of robust photometric redshift estimates (and their uncertainties) optimized for a range of science cases; software for characterizing the selection functions associated with Rubin data; support for time series analyses including the development of new light curve classification algorithms and the ability to scale these algorithms to the volume of data from Rubin; the ability to reprocess the Rubin images with algorithms optimized for specific science (e.g., detecting low surface brightness features in images); and custom generation of postage stamp images for sources detected by Rubin and other telescopes and the application of novel analysis algorithms to these postage stamp images.

The strengths and shortcomings of existing systems need to be understood in order to develop concrete plans to rectify gaps in these areas. Awareness of developments within

Rubin is important to enhance and not duplicate tools in the making. Enabling interoperation of multiple data sets either by co-location or by building effective streaming tools is an area of interest where IDACs could take the lead.

Development and implementation of these technical and science applications will require continued engagement around common software infrastructure by the science community. An inclusive approach to this work will need to be developed that addresses the needs of the science community as a whole (Section 4) and that employs best practices for fostering and maintaining engagement with those members of the research community who are traditionally underrepresented, whether as marginalized communities or as smaller or minority serving institutions. In Section 5, we describe how LINCC Frameworks will facilitate additional community discussions aimed at fostering broad and inclusive collaboration on cross-cutting software components that enable a large number of early science cases, building upon the technical areas identified in Section 3 and using the strategies and best practices from Section 4.

Contents

1. INTRODUCTION

The LSST, which will be carried out by the Vera C. Rubin Observatory (henceforth "Rubin Observatory" or "Rubin"), is the flagship ground-based astronomical survey of the 2020s. Every night, LSST will process 20 TB of images and deliver a stream of 10 million alerts for transient, variable, and moving objects in the sky. Each year, Rubin Observatory will reprocess all LSST images with state-of-the-art image processing software to build improved and deeper composite images of the southern sky, detecting tens of billions of objects and characterizing their properties. The scientific reach of the LSST will be extraordinary, addressing questions about the makeup of the Universe as fundamental as: how did the Solar System form; what processes govern the birth and death of stars; how does the dark matter in the Universe sculpt the shape of our own Galaxy; what is the nature of the dark energy that drives the accelerated expansion of our Universe?

In order to take advantage of this once-in-a-generation opportunity to transform our knowledge of the Universe, the astronomical community will need access to state-of-the-art analysis techniques that work at the scale and complexity of the LSST data. LSST's data products will include catalogs of sources, calibrated images, and a science platform to access and analyze these data. These resources must be supplemented with key algorithms and code that can enable petabyte-scale analyses to search for one-in-a-million or one-in-a-billion events in continuous streams of data, to identify trends and features within billions of sources, and to undertake sophisticated population analyses of the astronomical data.

These computational challenges are not generally unique to single areas of astrophysics. For example, the challenge of analyzing large samples of time series data are common, whether measuring periods of RR Lyrae stars to study the 3D distribution of stars in our Galaxy or classifying Type Ia supernovae to estimate their distance. Each requires access to multi-band time series data, the ability to run bespoke analyses on these data, and the ability to store and share the results of these analyses. If we can address these challenges as a community, we can improve and advance the science from Rubin data as a whole.

Equitable collaborations are as important for the progress of good science. This requires facilitated conversations on how to foster good communication, trust, and inclusion among research teams. By centering topics for equitable collaborations early in the definition of scientific collaborations, we can lay a foundation for building mutually beneficial partnerships with groups traditionally marginalized from the research process and create team environments that leverage diversity for scientific productivity.

The LINCC Frameworks initiative is a program supported by the Schmidt Futures Foundation to develop science-focused astronomical software that can enable a broad range of research projects. The goal of LINCC Frameworks is to accelerate scientific research through a combination of developing productionized, reusable analysis frameworks and to further inclusive collaboration through outreach and education. The LINCC Frameworks team will be responsible for identifying and enabling common computational needs within the science use cases, providing tools for adoption by the LSST SCs and other groups to use in the course of their analyses.

This whitepaper represents the results of the LINCC Frameworks "Data to Software to Science" meeting held at the Center for Computational Astrophysics at the Flatiron Institute between March 28th and March 30th 2022. In the context of LINCC Frameworks, the motivation for this whitepaper is to bring together a diverse group of researchers to discuss scientific use cases that could be undertaken with data from the first two years of the LSST survey, and to identify the common computational challenges in undertaking this research. Workshop participants were selected across a broad range of science interests, career stages, and demographics, and to represent researchers at large and small institutions together with those institutions that are historically underrepresented in astronomy. In this paper we discuss the science cases (section 2) and the technical challenges (section 3) that emerged as we discussed how to undertake early Rubin science. We describe best practises when establishing research collaborations that comprise a broad and diverse community in section 4. The detailed science use cases presented at the meeting are described in Appendix B and the technical use cases in Appendix C. The technical themes that were systematized from the science use cases and plans to develop these themes into scientific software development are outlined in section 5.

This is the first step in a process to identify and develop scientifically cross-cutting analysis software infrastructure to meet the science needs of Rubin, augmenting existing efforts within Rubin Observatory and the LSST SCs. The needs identified in this white paper will be further developed into software requirements and design documents in collaboration with the broader Rubin community.

2. MOTIVATING SCIENCE

Rubin Observatory and LSST were designed around four main science pillars (Ivezić et al. 2019) covering fairly broad areas. The impact of LSST will, however, extend well beyond the topics for which it was designed, enabling the scientific activities of a broad community. Currently, the primary organizational units within that community are the LSST SCs[1], which operate independently from each other but are connected within a federation via a charter. The LSST SCs vary in scope and size, have membership drawn from the international LSST science community, and when taken together have organized efforts to prepare for many of the scientific applications of LSST. Some SCs have written science roadmaps outlining their planned activities, and the steps to achieve them, using the data releases and other products (e.g., alerts) from Rubin Observatory.

The "Data to Software to Science" workshop identified seven primary scientific areas, defined based on expected commonalities. These were not explicitly tied to the LSST SCs – e.g., some SCs have an interest in multiple of the seven scientific areas, and some areas relate to the interests of several SCs. Organizers ensured representation at the workshop from all LSST SCs. Participants worked within the seven science areas to flesh out the high priority and high urgency analyses they would do within their SC with early LSST data. Some science use cases could fit into more than one of these seven areas, as they are not fully distinct; in that case, participants made an arbitrary choice of where to discuss the use case. A template was provided to guide participants in defining the steps needed for their analysis, while also identifying the associated computational needs and required software tools (Section A.1). The goal was to enable the identification of common software needs, which might foster collaborations between scientific communities on common infrastructure and could drive LINCC development priorities (i.e., to provide tools that those SCs would use in their analyses).

The use cases were not intended to be comprehensive, but rather to be examples of research that was of interest to the workshop participants. Many interesting science analyses may, therefore, not be represented in this whitepaper. Given the diversity of the science cases described in Appendix B we believe, however, that the technical cases that arose from our evaluation of the science cases is reflective of the computational and software challenges that the community will face in delivering on the potential of LSST. The use cases are compiled in Appendix B; below we outline some of the main themes that emerged.

Cosmology: Use cases in cosmology (Section B.8) include both static (e.g., cosmological weak lensing, clustering, or galaxy cluster analysis) and transient (e.g., Type Ia supernovae) science. Some key use cases also incorporate other survey data, such as from the Cosmic Microwave Background (CMB), or targeted follow-up observations (e.g., for supernova cosmology). For static use cases, key software infrastructure needs include photometric redshifts and associated tools, catalog-level cross-matching capabilities, and the ability to quantify the survey selection function. For transient science cases, time series analysis

[1] https://www.lsstcorporation.org/science-collaborations

(light curve representation and classification algorithms) is crucial. In a subset of cases, image reprocessing may be needed.

Extragalactic static: Use cases in this area (Section B.2) span a range of applications, from extragalactic stellar streams to dwarf galaxy populations, galaxy morphology and physical parameter studies, and galaxy cluster science. Key enabling technology includes photometric redshifts, custom postage stamp processing (e.g., to apply custom sky subtraction or deblending algorithms), cross-matching capabilities against high-resolution or multi-wavelength datasets, and software to track the selection function and observational effects across the survey. In a subset of cases, the ability to apply custom analysis or classification software is also needed. For example, solely morphological object classification is likely to be insufficient to discriminate between very distant galaxies and nearby stars of similar colors (brown dwarfs).

Extragalactic transients: Extragalactic transient science spans a range of use cases (Section B.3), from those that will use LSST alone or LSST in conjunction with follow-up data (based on events to which classifiers have been applied), to those that are more explicitly cross-observatory, such as the follow-up of detected gravitational wave events. Detection of transients relies crucially on the template creation process used to produce difference images, and on the ability to effectively represent light curves to enable time series analysis such as classification algorithms. Many use cases also rely on cross-matching capabilities and photometric redshifts. Depending on how templates are produced, additional image reprocessing capabilities may be needed.

Extragalactic variables: These use cases (Section B.4) typically relate to science with Active Galactic Nuclei (AGN), including strong gravitational lensing of quasars. Among the challenges are that variability is across a wide range of time scales, resulting in a need to analyze light curves over long time baselines. The optimal methodology to quantify the variability (e.g., structure functions) using the light curves is still to be determined. The ability to recognize variability is important to the triggering of follow-up observations (e.g., spectroscopy, and high-cadence photometry). Some AGN science also relies on observations other than optical ones, requiring cross-matching capabilities (along with cross-calibration) to include data from other observatories. Photometric redshifts for AGN pose challenges since some template-based photometric redshift methods may not include AGN among their templates.

Local Universe transient and variable: Science use cases in this category (Section B.6) relate to variable stars, and physical phenomena such as microlensing and binary systems. Because these are all non-static phenomena, efficient light curve analysis technology is essential to enabling these use cases. Many of the use cases require catalog-level cross-matching against multiwavelength datasets of a variety of sizes, from small (e.g., Gaia) to large (e.g., Roman Space Telescope), and some require a good understanding of selection functions.

Local Universe static: Science use cases in this area (Section B.5) range from understanding static phenomena in the Milky Way (e.g., stellar science), probing Milky Way

structure, and understanding populations of local dwarf galaxies and other aspects of the Local Group. The key enabling technologies identified across these use cases included catalog-level cross-matching, and software to understand the survey selection function. Image reprocessing and image-based analysis or classification algorithms, primarily used for dwarf galaxy and stellar stream identification, were identified as less common needs. Some use cases involving understanding Milky Way (MW) structure might rely as well on variable stars and hence have the same dependencies as the Local Universe transient & variable science cases.

Solar system: The solar system use cases (Section B.7) primarily focus on the detection and characterization of populations of objects within our solar system, though a subset also touch on solar system science outside of our own solar system (e.g., interstellar populations ejected from other solar systems). These use cases often involve use of both catalogs and images for the analysis, resulting in a need for large-scale image access (i.e., bulk reprocessing of calibrated images and the analysis of cutouts). The largest computational challenges arise in use cases that require additional image analysis and catalog/time-series post-processing. Data and software for computing selection functions is also needed.

The identified cross-cutting software needs from these areas are summarized in Section 3.

3. CROSS-CUTTING SOFTWARE AND INFRASTRUCTURE TO ENABLE EARLY SCIENCE

Though highly scientifically diverse, the science areas described in the previous section revealed a significant degree of commonality in terms of technical needs. This presents a significant opportunity: a development of a small set of components or services has the potential to enable a large number of early science cases. We identified seven distinct technical need areas:

1. Scalable Cross-matching
2. Photometric redshifts
3. Selection function determination
4. Time series analysis support infrastructure
5. Sky image access and reprocessing at scale
6. Object image access and analysis at scale, and the need for
7. Scalable job execution systems.

All but the last of these were discussed within scientifically-diverse breakout groups to better understand developments in each technical area that would support the various science use cases[2], with the identified connections between scientific and technical areas shown in Table 1.

	Cross-matching	Photo-z	Selection functions	Time series	Image reprocessing	Image analysis
Cosmology	✓✓	✓✓	✓✓	✓✓	✓	✓
Extragalactic static	✓✓	✓✓	✓✓		✓✓	✓
Extragalactic transient	✓✓	✓✓	✓	✓✓	✓	✓
Extragalactic variable	✓✓	✓	✓	✓✓	✓	✓
Local Universe transient & variable	✓✓		✓	✓✓		
Local Universe static	✓✓		✓✓		✓	✓
Solar system	✓		✓✓	✓✓	✓	✓✓

Table 1. Table highlighting the connection between scientific and technical areas discussed at the workshop. Rows are science areas while columns are for infrastructure capabilities. A double checkmark (✓✓) signifies that some infrastructure capability is essential to enable a particular scientific area, while a single checkmark (✓) signifies that the infrastructure capability would enhance or expand scientific discovery within that area but is not necessary to enable all of it.

The summary of each technical area is as follows:

Scalable Cross-matching: Nearly all science use cases presented at the Workshop require the ability to (generally positionally) cross-correlate the detections in the LSST catalog with one or more other catalogs – an operation commonly known as "cross-matching".

[2] Discussion of the scalable job execution system was postponed for a later date because not enough technical stakeholders in this area were present.

This capability would enable enrichment of LSST data with information taken in other wavelengths, at other times, different resolutions, or in general different characteristics. The capability is needed in two regimes:

1. Real-time – low-latency matching of $O(10\text{k})$ sources to $O(10)$ catalogs each with O(1Bn) objects (to support adding information to alert streams from other catalogs), and for
2. Offline data analytics – the ability to match O(10Bn) x O(1Bn) object catalogs, followed by joining data from both catalogs (e.g., full time series of observations, multi-wavelength studies) for analysis at scale.

In both regimes, the cross-matching capability **must** be easy to use for the end-user. For example, it may be provided at the community broker level for real-time cases, or accessible as simple Python (e.g. Pandas-like) calls or SQL-like statements callable from Python notebooks for the offline-level.

Importantly, cross-matching is *generally just the first step in a longer analysis process*, nearly always followed by fetching additional data from catalogs being cross-matched. Therefore, this operation is *better thought of as a (distributed) join of (large) tables on a spatial (user-defined) index, followed by (potentially heavy) computation on the result of the join*. This argues for the cross-matching capability to either be a part or seamlessly interface to a larger scalable analytics system.

This may also have implication for computation and storage service providers (facilities). For example, more emphasis may be needed on co-locating large datasets, enabling caching of frequently used remote datasets, as well as possible hardware considerations to enable efficient I/O.

Photometric redshift determination: A photometric redshift, or *photo-z*, is an estimate of the redshift (or distance) of an object made from available (generally multi-band) photometric information. Photo-z estimation is an algorithmically and computationally complex problem, and one that – depending on the science case – may not have a uniquely optimal solution. In some cases, photo-z estimation may incorporate some elements of classification (e.g., to distinguish between different extragalactic sources based on their SED). There was consensus among participants across all science areas about the need to move beyond simple errors in quantifying photo-z vs. spectroscopic redshifts, towards metrics targeting science cases (Section C.7.3). Currently, different scientific communities typically develop such metrics independently and apply them to specific photo-z codes in common use; in the era of LSST, shared infrastructure permitting application of different science-driven metrics to the outputs of multiple photo-z codes may be essential input to Rubin Observatory's choice of a single photo-z estimator to apply to a given data release, and could also help drive the scientific community's work on developing improved photo-z algorithms.

From a computational capability perspective, photo-z estimation codes need to be able to i) run on $O(5\text{Bn})$ galaxies at least as often as every data release, ii) allow for nightly calculation using multi-wavelength data for the Broker-filtered extragalactic alerts in any

given night, and iii) be able to update their estimates if additional information becomes available (e.g., spectra). Photo-z codes are expected to be highly data-parallel. The outputs are expected to be on order of $O(100)$ numbers per object (roughly $O(10\text{TB})$ for an LSST-scale dataset). Depending on the algorithm, they may be data intensive in terms of input (e.g., if they require cutouts of individual objects, and not just photometry). Significant prior art already exists, including codes such as RAIL and qp. More detail can be found in Section C.7.2.

While numerous science cases require photo-z estimates, we recognize they do so at varying degrees of complexity – from full posterior $p(z)$ PDFs, to simple point estimates. For many science cases, we would like to know things like i) Is there a single "peak" in the p(z), or is it multimodal?, ii) What is the "best-fit" redshift or the peak redshifts in case of multimodality? iii) What is the spread (2nd and higher moments)? or iv) Is there any other statistical property of the p(z) that correlates with object classification? Existing photo-z codes do not have a standardized way of reporting these things; we encourage some of this standardization to occur. We discuss these approaches in Section C.7.1.

Selection Functions: Selection functions are core components of any modeling procedure that aims to quantify the population statistics or density distribution of sources or objects. A selection function for a given modeling method may contain things like the detection efficiency of sources with a given brightness or shape, the classification accuracy of sources, the cadence of observations, the Milky Way and intergalactic dust distribution, or the crowdedness (in source counts) of a field.

The Rubin Data Management (DM) system will provide the core selection function data product: the detection efficiency of an ideal point source given as a function of position within the focal plane for each visit (direct and differenced), and the same quantity computed for each delivered coadd.

However, each specific science case, classification, or detection algorithm will need a specialized selection function which depends on the science question or model being studied. An example may be the detectability of objects within the Solar System as a function of their orbital and physical parameters, or the detectability of galaxies as a function of their morphology or surface brightness.

Building these will require additional software, and may require external catalog data or additional (potentially even pixel-level) processing. The detailed requirements are not clear at this point: we therefore recommend to engage the community in constructing worked examples of how to build and use selection functions of varied complexity for different use cases. We discuss some known options in more detail in Section C.3.

Time Series analysis support infrastructure: The LSST will deliver a uniquely large volume of high-quality multi-epoch measurements for each of the $O(40\text{ Bn})$ objects it is expected to detect. Converting this rich dataset into a wealth of science results will depend

of software infrastructure to i) extract features and classify the captured time-series, ii) enable parametric fitting, and iii) enable anomaly detection.

Variable sky astrophysics is a vast field requiring many period-finding and feature-identifying tools. Thus, to efficiently identify all sources of variability in (nearly) real-time for classification and catalog-creation, a complex multistage algorithm is required. Currently, users must make use of many other tools to handle these different types of variability, resulting in running many similar analyses on the same data set to tease out different features. And while many well-built and useful tools are available for feature extraction and classification, running them efficiently on LSST-scale datasets is challenging. We therefore explore the creation of a classification algorithm designed to handle the entirety of the variable-sky database generated by LSST, in combination with other southern-sky surveys. This tool would quickly and automatically classify transients and variables based on features in a multi-band light curve (shape, period, filter-specific amplitude and decay, etc) and the available color, magnitude, parallax, angular diameter information from LSST. The details are discussed in Section C.4.3.

Beyond classification, many analyses involve fitting a pre-specified model to data where the model parameters have semantic content (Section C.4.1). Given the numerous and diverse objects that Rubin LSST will observe the models which are fit to the observations must be flexible while also including physical information about each object. Commonly used tools in parametric time series analysis should be automatically computed (or trained) on all objects on a regular basis, along with specific subsets being analyzed with targeted tools as needed; all provided through a unified interface with a common data structure. Additionally, model selection criteria such as AIC or Bayes factors shall also be pre-computed to enable comparison between models, along with uncertainties and ranges of validity for model parameters. Providing the data alongside informative statistics in a networked and unified interface should help maximize the potential of LSST.

Finally, there will be some objects that are difficult to classify or otherwise unusual – *anomalies*. We discuss tools to detect them in Section C.4.2, primarily concentrating on real-time detection in LSST alert stream. To optimize the use of Rubin Observatory as a discovery machine for rare and high-priority events, infrastructure must be developed to efficiently identify anomalies among massive datasets in a timely manner. This will require synergy between machine learning tools for anomaly detection and visualization techniques for interactive and low-latency high-level analysis. These proposed tools should be sufficiently scalable and fast enough to enable prioritization and follow-up of rapidly-evolving events before they dim (early SuperNovae (SN) interaction, cometary outbursts or breakup, rapidly changing AGN, microlensing events/Tidal Disruption Events (TDEs)/KNe/other unknown phenomena and extreme cases of known types).

Sky Image Access and Reprocessing at Scale: While much of the science that will be done with LSST will rely entirely on the catalogs delivered by the science pipelines, a significant number of science use cases will also require analysis of the image data. This includes a

need for reprocessing of subsets ranging from cutouts of individual objects (covered next in this section) to larger image cutouts, full-focal plane data, or survey-scale pixel-level reprocessing as discussed in Section C.5 and summarized here.

The analysis of presented use cases point to the need for scalable i) data access services ii) processing infrastructure, and iii) processing software. For all these, the defining factor is the scale of the problem: $O(100\text{TB+})$ to $O(10\text{PB+})$ of image data. Supporting these use cases will require close collaboration between the Rubin Operations team (making sure the data is available in practice), infrastructure providers (whether HPC facilities, IDACs, or the public Cloud), and the users wishing to execute large-scale campaigns and write custom processing software.

We especially stress the importance of the first one of these, *performant data access*: services to quickly – O(seconds) – deliver a large number of custom-sized or -shaped cutouts up to the full dataset are critical to enabling this technological element. See Section C.5 for detailed discussion.

Object Image Access and Analysis at Scale: Finally, we look at the kinds of software tools that are likely needed to enable image-based analyses done at the level of *individual objects*. While there is overlap with the previous use case ('Sky Image Access and Reprocessing at Scale'), we discussed this use case separately as it is both scientifically different – we assume the positions of objects are known – but also as it may be technically different: here the access patterns are fundamentally *object-aligned*, and may benefit from a different storage strategy to be made performant.

A core element for this type of analysis is an (object-level) scalable image cutout service. This service must be able to i) make image cutouts of any input image over scales ranging from $\sim$ 10arcsec to $\sim$ 10arcmin, ii) optionally deblend objects in the scene, iii) provide the ability to link the items on the resulting image to archival data (both catalogs and images) from external sources, and iv) allow for filtering of selected cutouts (e.g., different bands). Ideally, to support alert-triggered follow-up, such a service would also be able to return results in real-time (order of seconds), at least for most recent imaging data. The details are discussed in Section C.6.

4. INCLUSIVE COLLABORATION

This section builds on discussions held as part of a breakout session at the LSSTC From Data to Software to Science Workshop. The session focused on having participants think about how to make more equitable collaborations. Participants were separated into small breakout groups and given use cases that touched on themes of the inclusivity of collaborations, specifically focused on issues of cross-institutional partnerships, the allocation of credit, resource access, and working with students. Each use case was paired with a set of questions that prompted participants to discuss challenges to the collaborative research process, as well as potential solutions to those challenges (see Appendix D for a full set of use cases and questions provided to breakout groups).

Below, we present key themes that arose during these breakout discussions, but rather than present all points that participants mentioned, we have instead chosen to highlight the themes that emerged that align with the social science literature on inclusive collaboration. First, we detail various challenges that face research collaboration, and then we outline solutions that could be implemented to ensure research collaborations are more inclusive. Many of the challenges and solutions below center on specific policies or practices, such as how to incorporate students more effectively into teams, how to build cross-institutional partnerships, or how to structure team environments at an organizational level in ways that will create more inclusive outcomes for all team members. These recommendations must be considered in the broader context of team diversity, specifically demographic diversity. Astronomy and astrophysics has become increasingly diverse along lines of race and gender in recent decades (Merner & Tyler 2017; Porter & Ivie 2019). Research on team diversity has demonstrated that more diverse teams are typically more creative, innovative, productive, and impactful than homogeneous teams (Herring 2009; Kalev 2009; Freeman & Huang 2014). Fully integrating underrepresented researchers in the research team allows for the level exchange of resources like funding, expertise, or information, and thus aids in equity in the field while also creating the potential for diversity to have positive outcomes for research teams. However, beyond the fact that incorporating more underrepresented researchers into astronomical research collaborations may make the outcomes of research more impactful, there is a moral responsibility to have research teams be more representative of the population regardless of their impact. This means more than just increasing the representation of underrepresented groups. Sociological research has demonstrated that too much focus on increasing representation can lead to dynamics of tokenism, which can mean potentially more opportunity but often leads to more work and harsher evaluations for underrepresented groups than their colleagues (e.g., Kanter 1977; Harvey Wingfield & Myles 2014; Alegria 2019). Instead, we have a responsibility to structure collaborative teams in ways that will fully integrate traditionally underrepresented scholars (Smith-Doerr et al. 2017).

In order to integrate underrepresented researchers into collaborative teams means we must actively work to organize our teams in ways that will allow these folks to share their experiences and expertise in meaningful ways. Others have noted that intentionally

structuring research collaborations is a key strategy for diversifying astronomy in the coming decade (Bechtol et al. 2019). It takes intentional efforts to design inclusive teams or enact authorship processes that will fully integrate underrepresented groups in meaningful ways. The recommendations put forth below are organizational strategies for how to better structure our research collaborations, with the intent of better incorporating the diversity of astronomy and astrophysics into the research process.

4.1. *Challenges in research collaborations*

Research collaborations have become increasingly common across Science, Technology, Engineering and Math (STEM) disciplines (Bozeman et al. 2013; Leahey 2016). While collaborations are often noted for producing more innovative and high-impact research (Frickel et al. 2016), collaborative research often faces several challenges that can undermine the efficacy of the project. For instance, collaborators from different types of institutions[3] (research, teaching) often face differing evaluative pressures around productivity or efficiency. University administrators at research institutions often primarily evaluate scientific faculty by the grants they secure, the prestige of the journals in which they publish, and the impact of their research programs. In contrast, administrators at teaching institutions likely expect faculty to juggle higher teaching loads. Astronomers at teaching institutions are evaluated by course evaluations and student mentorship, but are also expected to publish, although it is often not weighed as heavily in tenure or promotion. Thus, these divergent expectations of astronomers from research and teaching institutions may lead to tensions in the collaborative process.

Time pressures are intertwined with evaluative pressures for promotion or tenure, especially for collaborators that are housed at different types of institutions. Past scholarship has shown time to be an important ingredient for successful collaborations as well as a consistent source of tension among collaborators (Parker & Hackett 2012; Sacco 2020). Teaching institution astronomers with higher teaching loads than astronomers at research institutions have their research times constrained by the rhythms of semester schedules. More teaching translates to less time for research and vice versa. The teaching loads at teaching institutions may also prohibit or severely restrict attendance at professional conferences; for instance, the winter American Astronomical Society (AAS) meetings held in January of each year are less accessible to astronomers who are tied to the semester schedule. Thus, expectations that collaborators will prioritize research or be able to disseminate findings at key conferences in the field often have astronomers at research institutions in mind. This can lead to challenges between astronomers from different types of higher education institutions.

Authorship credit is consistently a source of tension in research collaborations (Tsai & Hsu 2014; Bozeman & Youtie 2017; Youtie & Bozeman 2014; Sacco 2020; Misra 2020). The uneven allocation of credit or unclear processes for allocating credit may contribute to a

[3] There is a wide variety of types of institutions, especially when considering the differing landscapes in different countries. Even just within the US, there are many types of institutes - research and teaching universities, government labs, minority-serving institutions, and more. However, for simplicity we use just two examples, research and teaching institutions, that already illustrate some of the challenges that can arise due to differing priorities and levels of research infrastructure available.

team environment that does not promote inclusion. Credit on publications is the coin of the realm in science; scholars at both teaching and research institutions are expected to publish to meet tenure and promotion milestones. In addition, students looking to be competitive on the job market after graduating are expected to publish as well. Research collaboration has become more common in recent years (Leahey 2016), which on one hand opens up more opportunities for scholars to publish, but also creates scenarios in which collaborators are competing for authorship status on the fruits of their collaboration. Tensions around authorship are particularly challenging for junior researchers, who rely on publications in order to move their careers forward (Tsai et al. 2016). It is common for senior or more well-known researchers to gain undue credit or be listed higher in authorship order than is often warranted (Merton et al. 1968; Zuckerman 1977; Rigney 2010; Tsai et al. 2016).

There are several factors that can undermine the inclusivity of a team. For instance, the reliance on technology to support collaborations that span multiple institutions may undermine inclusion for some team members. Phone calls, Zoom, Slack, or other technologies that mediate non-face-to-face interaction between team members makes collaborations less personal. Research on the effects of virtual technology on work have found that these kinds of technologies can make individuals feel alienated from their team, and also cause some individuals to self-censor their opinions in ways they may not in face-to-face interaction (Soga et al. 2021). Virtual technology can make remote teamwork especially hard for women and other underrepresented minorities in the workplace (Bolade-Ogunfodun et al. 2022).

Other organizational dynamics of collaborative teams can undermine the inclusivity of a team environment as well. For instance, team inclusivity may be undermined as a result of when or where a meeting is held. Team meetings may be scheduled at times that are more convenient for some members than others. This may be especially true when there is a small group of leaders making decisions for the team, as a concentrated hierarchy is less likely to include the views of all team members. In addition, holding team meetings in physical locations like bars or holding meetings "after hours" may be exclusive to the lifestyles of some team members.

Another challenge is the incorporation of students onto collaborative projects. Scholars at research and teaching institutions also likely have differing access to student support for their research. Many research institution scholars have access to postdocs and graduate students (as well as interested undergraduates). The technical expertise and mentorship provided by postdocs and graduate students facilitates student support of research institution scholars' research. In contrast, teaching institution astronomers may be working primarily with undergraduate students who have limited expertise and who require additional mentoring and instruction to carry out research.

4.2. *Recommendations*

4.2.1. *Building sustainable cross-institutional partnerships*

In building sustainable cross-institutional partnerships, collaborative teams must work to actively manage expectations of what the team will accomplish and who will take on specific tasks. It is important for all collaborators to outline their expectations for the partnership before it is well underway as a way to avoid future conflicts. For instance, it would be fruitful for all team members to clearly define how they envision their role in the project as well as what they are hoping to get out of the collaboration. Each team member should outline their expectations for their workload, such as what their collaborators can expect of them, what a feasible workload looks like based on their other professional commitments, and task breakdowns. Team members should also lay out what they see as a realistic timeline for research in relation to their teaching load and other pressures. This is especially important in collaborations that span different types of institutions, as research- and teaching-institution scholars face different evaluative pressures from their administrations. Cross-institutional partnerships that include both research- and teaching-institution scholars should consider the different ways in which semester schedules shape team members' workloads.

In addition to laying out general expectations at the start of the project, team members should outline expectations for producing publications, the allocation of credit, and other types of dissemination. By allocation of credit, we refer to the arrangement of authorship on publications produced by the collaboration. Cross-institutional partnerships will benefit from detailing a clear division of labor upfront and articulating how this will translate to credit on publications or presentations. It is especially important to have these conversations upfront when a cross-institutional partnership spans both research- and teaching-institutions. Scholars should be assigned tasks that complement their strengths, but also what is valued by their institutions. For teaching-institution scholars, this may include identifying positions that are more directly tied to teaching or outreach (if that is what the administrators at their institution value). This may also include dissemination of research findings in an array of journals or conferences that will be recognized as beneficial for the varying members of the team.

All members of a cross-institutional partnership should be provided with the resources necessary for their success. This may be funding or computational resources that are necessary for their science to succeed. This could also be an exchange of personnel that would benefit the collaboration as well as the careers of those being exchanged. For instance, undergraduates from teaching institutions could be sent to work with research-institution scholars for closer mentorship from faculty members or graduate students as a way to prepare for the research process. Similarly, graduate students or postdocs housed at a research institution could go work with a partnering teaching institution as a way to gain experience with teaching, mentoring, or outreach.

It is important for cross-institutional partners to maintain some degree of flexibility as well. Often, research does not go according to plan, the person who initially volunteered to draft a paper may no longer have time, or unforeseen institutional pressures may change a team members' commitment to the project. Cross-institutional partnerships should periodically reassess priorities and goals of the project. This will allow collaborators to raise any issues

or concerns they have, allow teaching institution astronomers to raise any unforeseen ways the collaboration conflicts with their other priorities, and allow the team to strategize ways to adapt to pressures or pitfalls as the project progress.

4.2.2. *Strategies for dealing with authorship*

Authorship and credit allocation is one of the most contentious aspects of research collaboration (Misra 2020; Sacco 2020). To mediate these tensions, collaborators must outline authorship expectations early on and make clear how credit will be allocated before they begin drafting papers. When determining authorship, it is often useful to clearly outline who contributed what to the paper, such as who developed an idea, who developed code, and who wrote what portions of the draft. It is also useful if all collaboration members agree in advance on a well-defined process for defining the authorship order. As a way to benefit junior researchers on a team, collaborators could potentially list junior people higher in authorship to benefit their careers. This is an option to be discussed by the team at the start of a collaboration. A team may also decide to write multiple papers so that everyone can benefit more equitably. Students and postdocs would benefit from designing mentoring plans with their advisors that train these junior scholars on co-authorship, and lay out expectations for authorship and publication.

4.2.3. *Strategies for fostering an inclusive team environment*

Several factors shape the inclusivity of a collaborative team environment. Some factors are difficult to control. For instance, in cross-institutional partnerships, it may be unrealistic to have the full team be able to meet face to face. Technology often must mediate these relationships, which can make things feel less tight knit with more people feeling detached. While virtual mediums like Slack, Zoom, or other technologies make connectivity easier, they are also less personal and can lead to certain team members feeling alienated.

There are several things that teams can do regardless of whether teams are meeting in person or virtually. At the most basic level, giving each member of the team a role or task to encourage buy-in will get people more invested in the outcome of a team. Team members (especially junior team members) should be given opportunities to share their ideas. Regular meetings are important to the success of collaborative teams, but not all team members necessarily feel comfortable sharing in large group discussions. Thus, team leaders should encourage feedback and meet with individual team members on a regular basis to solicit input that members may not feel comfortable sharing in the large group. This will ensure that the diversity of voices on a team are incorporated into the decision making of team leadership.

Team members should also be given the opportunity to share work in progress to get feedback from collaborators. This can mean setting aside times during regularly scheduled meetings, or setting aside blocks of time specifically for “chalk talks” or work-in-progress talks. These opportunities to share work should specifically feature diverse members of the team. Providing opportunities for team members to share work in progress can benefit

research by generating more collaborative opportunities, as well as make varying members of the team feel recognized as researchers (Misra 2020).

Structurally, collaborative teams can improve the inclusivity of their environment by putting into place transparent processes for team decision making, like blind voting with written ballots rather than giving consent during team meetings. Teams should also work to improve by soliciting open and honest feedback from team members. To do this, teams should put into place structures for team members to critique aspects of the project they feel could improve without fear of retaliation. Teams should also enact feedback mechanisms to solicit input from all collaborators. Having all members read and agree to a collective code of conduct is another strategy for fostering an inclusive environment. Regularly communicating with one another over research successes is a third. For large collaborations, creating a shared team calendar can help team members keep track of meetings, and creates a shared norm around prioritizing team events.

Scholarship on collaboration shows that getting together to work face-to-face enhances the productivity of a team (Parker & Hackett 2012). Social events both on campus as well as other places (grabbing coffee as a team, going out to dinner, likely at conferences for cross-institutional collaborators) help connect team members to the broader collaboration and fosters a sense of community among the team.

4.2.4. *Strategies for working with students*

Collaborations face some unique challenges in incorporating students like undergraduates or early graduate students onto projects. As a result, collaborative teams should work to foster an inclusive environment in which students feel comfortable asking questions and proposing ideas without risk of being ridiculed by more senior members of the collaborative team. Virtual technology like Slack or one-on-one meetings with research team faculty can create opportunities for students to pose questions that they may feel uncomfortable asking in large team meetings. Another component of fostering an inclusive environment for students is being mindful of student familiarity with different aspects of the research process in team discussion. Students likely have less technical experience or language proficiency than more senior researchers, and a lack of mindfulness to this fact can create barriers to student participation or success in the process. Teams should also be mindful of student availability for research, and how the collaboration may come into conflict with other work or academic commitments they may have.

One potential strategy for incorporating undergraduates or early graduate students within collaborative teams is giving them tasks that will set them up for success. This could be defining small tasks that are appropriate for novices to undertake, that will not be too challenging but also help encourage confidence in their research abilities. Similarly, some attendees proposed creating a hierarchy of tasks, ranging from easy to more difficult, as a strategy for incorporating students into the research process progressively. As such, faculty members in the collaboration should be prepared to mentor more junior students based on the difficulty of the tasks at hand. Teams could even develop starter projects for students

(building starter projects into the design of research proposals) with novice students in mind, rather than dropping them into the deep end of astronomical or astrophysics research.

Collaborative teams looking to work with undergraduate or early-career graduate students need to be mindful of how the research will benefit them. For instance, setting up structures that will help students develop skills, such as data analysis, coding and scientific writing, would be beneficial. Working with students to have them assess their strengths, weaknesses and goals within the research environment can foster a sense of ownership. Creating opportunities for students to expand their social networks is also important. Students should not be viewed as free or cheap labor. Part of this is ensuring that students will be credited for their work on a project, both in publications and in public presentations of research findings.

4.3. *Conclusion*

While astronomy and astrophysics have become increasingly diverse in recent decades, more efforts need to be made to integrate traditionally underrepresented researchers into the research process. We view the organizational structure of research collaborations to be a fruitful area for interventions that will ultimately make astronomy and astrophysics more inclusive and equitable in the coming decade. In this white paper, we have outlined what we see as some core challenges to structuring inclusive collaborations in astronomical research, as well as recommendations for addressing these challenges.

5. NEXT STEPS

We begin with some top-level proposals for next steps after this workshop, then discuss follow-up activities in specific technical or domain areas.

5.1. *Top-level proposals for next steps*

As demonstrated in previous sections, the community would benefit from continued engagement on common software infrastructure needs that would enable them to effectively carry out their high-impact scientific analyses with LSST data. Many software infrastructure areas would serve multiple scientific stakeholders, which is a point in favor of coordination across existing groups in the Rubin Observatory LSST scientific community, including Rubin Observatory, the LSST SCs, LINCC Frameworks, and teams developing IDACs. This coordination will promote effective use of the resources available to the community, while respecting the differing goals and scopes of the above-mentioned groups.

As a rule, for each area of effort identified, we envision near-term and longer-term activities. In the near term, a reasonable approach would be for the interested stakeholders (identified below in Section 5.2) to discuss the needs, elaborating on the initial understanding outlined in this white paper, and outline a pathway to determining how best to meet them. Depending on the state of existing tools in this area, that could include steps to determine the performance of the existing tools and opportunities to improve them to meet the needs of the LSST science community, or it could involve sketching out the design for a new tool if improving an existing one does not seem like a viable pathway forward. In several of the technical areas, existing software may do some of what is needed, but scalability is an issue. Working demonstrations of these analyses on existing datasets can be used to identify the computational tall poles and potential avenues for optimization. Identifying responsible parties with the resources to develop these tools, and any formal collaboration development, should be another part of those discussions, setting the stage for the longer-term phase (in which the work is carried out).

A first opportunity at starting those discussions is the 2022 Rubin Project & Community Workshop, where representatives from all of the stakeholders will be present. The LINCC Frameworks team, in particular, has a session with the goal of presenting the use cases developed here and identifying areas for progress. There are also sessions aimed at matching use cases and communities to IDAC resources, and specific sessions organized by the LSST SCs, such that there will be several opportunities for productive discussions. Depending on the outcomes of these discussions, there will be additional opportunities for virtual discussion outside of and after this meeting.

In several of the areas for future work described herein, ongoing work is taking place within Rubin Observatory, LSST SCs, or other groups, with a variety of collaboration norms and publication policies. These groups are encouraged to review the recommendations for inclusive collaboration in Section 4.2 and identify opportunities to adopt best practices described therein, which could be particularly important if collaborations form across groups. For example, explicitly writing down expectations for resource allocation,

publication of results and authorship of resulting publications, and mechanisms to permit full collaborative engagement at all career stages would be highly valuable to set the stage for those collaborations.

Finally, we note that some of the common infrastructure needs identified across science use cases can be satisfied purely through software development. Others may motivate rethinking which datasets are hosted at major computing centers, and in what form, in addition to motivating software infrastructure development. For example, catalog-level cross-matching needs fall into the latter category, and may deserve particular attention when considering the development of IDACs and other settings where users anticipate carrying out early LSST science analyses.

5.2. *Follow-up activities in specific technical or scientific areas*

In this section, we highlight more specific follow-up activities, which could follow the pattern described in Section 5.1 of including a discussion amongst key stakeholders (and open to all interested members of the LSST science community) to outline a plan for assessing how existing or newly developed tools might meet the identified needs. More concretely, a list of anticipated steps is as follows, though in practice some variation may be needed for each area.

1. Identify the key stakeholders and ensure they (and the broader community) have an opportunity to join the discussion.
2. Work with domain experts to refine the requirements, starting from the use cases and requirements outlined in this document, but potentially including others as well.
3. Identify existing software that could fill some of the needs, potentially with further development (again, in some cases this has already been done within this white paper). If there is any, then next steps could involve carrying out analysis of precursor survey data to benchmark performance and understand the key challenges in extending the software to work at LSST scale and precision.
4. Depending on the outcome of the previous exercise, develop a plan for next steps, whether it involves modifying existing software or developing new software.
5. Circulate to the community via https://community.lsst.org/ and communication channels used by the groups involved (e.g., Rubin DM, LSST SCs, LINCC Frameworks, etc.), to get feedback and further discussion.
6. Secure resources and formalize collaborations needed to carry out the work.

Table 1 highlights the connections between scientific areas and technical infrastructure, which can be used to identify relevant LSST SCs that should be consulted for each technical area.

There are several cross cutting technologies that need future exploration with close coordination amongst members of Rubin Observatory DM, IDACs teams, and stakeholders from the science community. Care needs to be taken to ensure any new development is complementary to the capabilities provided by and compatible with Rubin DM. These cross cutting technologies include batch processing (as mentioned in Section 3 this was not discussed

in depth at the workshop, but is an important area), image reprocessing, and image-based analysis. For example with image-based analysis, a natural step would be to assess whether the image cutout services provided with the RSP meet the needs described for image-based analyses at the level of individual objects, based on this white paper.

Communities working on extragalactic science and cosmology should identify opportunities for joint photometric redshift infrastructure, especially infrastructure for uncertainty quantification and science-driven metric development. Given that Rubin Observatory will provide the results of running a single photometric redshift algorithm in each release[4], the existence of a community testbed to enable further algorithm development and feed the algorithms back to Rubin Observatory for future releases (along with information about impact across the science community) will be important. Work within the LSST Dark Energy Science Collaboration (DESC) on Redshift Assessment Infrastructure Layers, https://github.com/LSSTDESC/RAIL (RAIL) may serve as the foundation for a community-wide photometric redshift testbed, and a first step may be to assess whether it can naturally be extended to the range of capabilities discussed in this white paper.

Selection functions is an area where the key functionality is expected to come from Rubin Observatory. Therefore, in this area we propose discussions should focus on what Rubin DM will provide, what additional functionality may be needed on top of that (based on worked examples), and what software capabilities may already exist that could fill that niche or be extended so as to fill it.

While the communities working on time domain science are currently distributed across nearly all LSST SCs, they should explore the potential for joint software infrastructure for light curve storage and analysis that would meet their needs while reducing the software development burden on any one group. Given the need for interoperability with the light curves provided by Rubin Observatory and by alert brokers, including those groups in the discussion is important. Improving existing light curve routines in Astropy until they work at the needed scale would be one pathway for exploration.

Regarding catalog-level cross-matching, this discussion has not only software implications (to support a variety of scales) but also affects how and where ancillary datasets are stored and made available to the LSST science community. As outlined in Section 3, cross-matching is important in different regimes for different types of analyses. A reasonable next step may be to delve deeper into these use cases and develop a set of requirements for those tools, along with some example precursor survey analyses with a variety of datasets that would help drive development.

The forms of collaborative work (who is involved, is it driven by many groups or few, etc.) will determine how the work is carried out. However, the potentially new collaborations to develop software infrastructure following the above actions should follow the recommendations in Section 4.2 as appropriate for the form of that collaboration. For example, this

[4] This will be chosen following rigorous selection between a number of candidates; for details cf. https://community.lsst.org/t/pz-lor-a-summary-of-the-proposed-pz-estimators-dm-shortlist/6308.

includes outlining roles and responsibilities, expectations for disseminating the work and allocating credit, and fostering an inclusive team environment.

ACKNOWLEDGMENTS

This workshop and the resulting white paper were initiated as part of the LINCC Frameworks program, which was supported through the generosity of Eric and Wendy Schmidt by recommendation of the Schmidt Futures program. We acknowledge workshop support from Heising-Simons Foundation award 2020-1916 to LSST Corporation (LSSTC). We thank the Center for Computational Astrophysics (CCA) at the Flatiron Institute (a part of the Simons Foundation) and especially the administrative staff (Kristen Camputaro and Fatima Fall) for hosting the "Data to Software to Science" workshop that provided the catalyst for this document.

C.O.C.: This material is based in part upon work supported by the National Science Foundation Graduate Research Fellowship Program under grant No. (2018258765). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. M.J.: This material is based upon work supported by the National Science Foundation under Grant No. AST-2003196. MJ, JM, HRS, SS, JRAD, and AJC wish to acknowledge the support from the University of Washington College of Arts and Sciences and the DiRAC Institute. The DiRAC Institute is supported through generous gifts from the Charles and Lisa Simonyi Fund for Arts and Sciences and the Washington Research Foundation. Y.-Y.M. was supported by NASA through the NASA Hubble Fellowship grant no. HST-HF2-51441.001 awarded by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Incorporated, under NASA contract NAS5-26555. AC acknowledges support from NSF awards AST-1715122, AST-2107800, and OAC-1739419 and DOE awards DE-SC0011665 JAVM aknowledges financial suport from CONACyT 252531. ABK acknowledges funding provided by University of Belgrade-Faculty of Mathematics (the contract 451-03-68/2022-14/200104), through the grants by the Ministry of Education, Science, and Technological Development of the Republic of Serbia. YT acknowledges the support of DFG priority program SPP 1992 "Exploring the Diversity of Extrasolar Planets" (TS 356/3-1). L. Č. P. is supported by the Ministry of Education, Science and Technical development of R. Serbia (Project No 451-03-68/2022-14/ 200002). CONACyT México under Grants No. 286897 and the Instituto Avanzado de Cosmología Collaboration. GF acknowledges the support of the European Research Council under the Marie Skłodowska Curie actions through the Individual Global Fellowship No. 892401 PiCOGAMBAS. Work by RS, RB, LV, and SU was supported by the Preparing for Astrophysics with LSST Program, funded by the Heising Simons Foundation through grant 2021-2975, and administered by Las Cumbres Observatory. RB acknowledges support from the project PRIN-INAF 2019 "Spectroscopically Tracing the Disk Dispersal Evolution". ĐS acknowledges the funding provided by Astronomical Observatory (the Ministry of Education, science and technological development of Republic Serbia contract 451-03-68/2022-14/ 200002). CSA: This work was partially enabled by funding from the UCL Cosmoparticle Initiative. AIM acknowledges support from the Max Planck Society and the Alexander von Humboldt Foundation in the framework of

the Max Planck-Humboldt Research Award endowed by the Federal Ministry of Education and Research. The Flatiron Institute is supported by the Simons Foundation. DI acknowledges funding provided by University of Belgrade-Faculty of Mathematics (the contract 451-03-68/2022-14/200104), through the grants by the Ministry of Education, Science, and Technological Development of the Republic of Serbia, and the support of the Alexander von Humboldt Foundation. XL acknowledges support of the National Science Foundation Grant No. 2108841: "Detecting and studying light echoes in the era of Rubin and Artificial Intelligence"; University of Delaware General University Research award GUR20A00782. AG acknowledges the financial support from the Slovenian Research Agency (grants P1-0031, I0-0033, J1-8136, J1-2460). S.D. is supported by NASA through Hubble Fellowship grant HST-HF2-51454.001-A awarded by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Incorporated, under NASA contract NAS5-26555. JLS acknowledges NSF award AST-1816100. T.A. acknowledges support from ANID-FONDECYT Regular 1190335, the Millennium Science Initiative ICN12_009 and the ANID BASAL project FB210003. KESF is supported by NSF AST-1831415 and Simons Foundation Grant 533845 and by the Center for Computational Astrophysics of the Flatiron Institute. LNdC would like to acknowledge the support from the Heising-Simons Foundation and the CNPq/FAPERJ program INCT do e-Universo.

Rubin Observatory is a joint initiative of the National Science Foundation (NSF) and the Department of Energy (DOE). Its primary mission is to carry out the Legacy Survey of Space and Time, providing an unprecedented data set for scientific research supported by both agencies. Rubin is operated jointly by NSF's NOIRLab and SLAC National Accelerator Laboratory (SLAC). NOIRLab is managed for NSF by the Association of Universities for Research in Astronomy (AURA) and SLAC is operated for DOE by Stanford University.

REFERENCES

Abbott, T. M. C., Aguena, M., Alarcon, A., et al. 2022, Phys. Rev. D, 105, 023520

Acero-Cuellar, T., Bianco, F., Dobler, G., Sako, M., & Qu, H. 2022, arXiv e-prints, arXiv:2203.07390

Akhlaghi, M. 2019, arXiv e-prints, arXiv:1909.11230

Akhlaghi, M., & Ichikawa, T. 2015, ApJS, 220, 1

Akras, S., Guzman-Ramirez, L., Leal-Ferreira, M. L., & Ramos-Larios, G. 2019a, ApJS, 240, 21

Akras, S., Leal-Ferreira, M. L., Guzman-Ramirez, L., & Ramos-Larios, G. 2019b, MNRAS, 483, 5077

Alegria, S. 2019, Gender & Society, 33, 722

Aleo, P. D., Malanchev, K. L., Pruzhinskaya, M. V., et al. 2022, NewA, 96, 101846

Allard, F., Hauschildt, P. H., Alexander, D. R., & Starrfield, S. 1997, ARA&A, 35, 137

Allen, D. A. 1984, PASA, 5, 369

Alonso, D., Sanchez, J., Slosar, A., & Collaboration, L. D. E. S. 2019, Monthly Notices of the Royal Astronomical Society, 484, 4127

Alves, C. S., Peiris, H. V., Lochner, M., et al. 2022, ApJS, 258, 23

Amon, A., Gruen, D., Troxel, M. A., et al. 2022, PhRvD, 105, 023514

Andreoni, I., Margutti, R., Salafia, O. S., et al. 2022, ApJS, 260, 18

Angus, R., Beane, A., Price-Whelan, A. M., et al. 2020, AJ, 160, 90

Asgari, M., Lin, C.-A., Joachimi, B., et al. 2021, A&A, 645, A104

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

Bachelet, E., Norbury, M., Bozza, V., & Street, R. 2017, AJ, 154, 203

Baldassare, V. F., Geha, M., & Greene, J. 2018, ApJ, 868, 152

—. 2020, ApJ, 896, 10

Bartos, I., & Kowalski, M. 2017, Multimessenger astronomy (IOP Publishing Bristol)

Beaumont, C., Goodman, A., & Greenfield, P. 2015, in Astronomical Society of the Pacific Conference Series, Vol. 495, Astronomical Data Analysis Software an Systems XXIV (ADASS XXIV), ed. A. R. Taylor & E. Rosolowsky, 101

Bechtol, E., Bechtol, K., BenZvi, S., et al. 2019, in Bulletin of the American Astronomical Society, Vol. 51, 216

Bechtol, K. 2017, SEARCHING FOR DWARF COMPANIONS OF THE MILKY WAY AND BEYOND IN THE LSST ERA

Becla, J. 2013, LSST Database Baseline Schema

Bellm, E. C. 2021, Review of Timeseries Features

Bellm, E. C., Kulkarni, S. R., Graham, M. J., et al. 2019, PASP, 131, 018002

Belokurov, V., Erkal, D., Evans, N. W., Koposov, S. E., & Deason, A. J. 2018, MNRAS, 478, 611

Berry, M. W., Mohamed, A., & Yap, B. W. 2019, Supervised and Unsupervised Learning for Data Science, 1st edn. (Springer Publishing Company, Incorporated)

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

Bianco, F. B., Ivezić, Ž., Jones, R. L., et al. 2021, arXiv e-prints, arXiv:2108.01683

Birrer, S., & Amara, A. 2018, Physics of the Dark Universe, 22, 189

Blackman, J. W., Beaulieu, J. P., Bennett, D. P., et al. 2021, Nature, 598, 272

Bochanski, J. J., Hawley, S. L., Covey, K. R., et al. 2010, The Astronomical Journal, 139, 2679

Bolade-Ogunfodun, Y., Soga, L., & Nasr, R. 2022, Rebalancing gender inequity and the digital divide: unintended consequences of working from home

Bolton, A., Eisenstein, D., Olsen, K., et al. 2018, https://docushare.lsstcorp.org/docushare/dsweb/Get/Document-30603/bolton_desi_overlap_mini.pdf

Bonito, R., Venuti, L., et al. 2021, https://docushare.lsst.org/docushare/dsweb/Get/Document-37625/rbonito.pdf

Bonito, R., Hartigan, P., Venuti, L., et al. 2018, arXiv e-prints, arXiv:1812.03135
Boone, K. 2019, The Astronomical Journal, 158, 257
—. 2021, The Astronomical Journal, 162, 275
Boquien, M., Burgarella, D., Roehlly, Y., et al. 2019, Astronomy & Astrophysics, 622, A103
Boubert, D., & Everall, A. 2022, MNRAS, 510, 4626
Bouy, H., Bertin, E., Moraux, E., et al. 2013, A&A, 554, A101
Boyajian, T. S., von Braun, K., van Belle, G., et al. 2012, ApJ, 757, 112
Boyajian, T. S., LaCourse, D. M., Rappaport, S. A., et al. 2016, MNRAS, 457, 3988
Bozeman, B., & Youtie, J. 2017, The Strength in Numbers: The New Science of Team Science
Bozeman, T. D., Scogin, S., & Stuessy, C. L. 2013, European Journal of Science Education, 17, 1
Bozza, V. 2010, MNRAS, 408, 2188
Brandt, W. N., Ni, Q., Yang, G., et al. 2018, arXiv e-prints, arXiv:1811.06542
Brescia, M., Salvato, M., Cavuoti, S., et al. 2019, MNRAS, 489, 663
Bricman, K., & Gomboc, A. 2020, The Astrophysical Journal, 890, 73
Broccia, G. 2021, arXiv e-prints, arXiv:2105.00089
Brown, W. R., Kilic, M., Kosakowski, A., et al. 2020, ApJ, 889, 49
Bullock, J. S., & Johnston, K. V. 2005, ApJ, 635, 931
Burdge, K. B., Prince, T. A., Fuller, J., et al. 2020, ApJ, 905, 32
Burke, C. J., Shen, Y., Blaes, O., et al. 2021a, Science, 373, 789
Burke, C. J., Liu, X., Shen, Y., et al. 2021b, arXiv e-prints, arXiv:2111.03079
Burke, D. L., Rykoff, E. S., Allam, S., et al. 2018, AJ, 155, 41
Burns, C. R., Stritzinger, M., Phillips, M., et al. 2010, The Astronomical Journal, 141, 19
Butler, N. R., & Bloom, J. S. 2011, AJ, 141, 93
Carlin, J. L., Mutlu-Pakdil, B., Crnojević, D., et al. 2021, ApJ, 909, 211
Carlsten, S. G., Beaton, R. L., Greco, J. P., & Greene, J. E. 2019, ApJ, 879, 13
Carrasco-Davis, R., Reyes, E., Valenzuela, C., et al. 2021, AJ, 162, 231
Chambers, K. C., Magnier, E. A., Metcalfe, N., et al. 2016, arXiv e-prints, arXiv:1612.05560
Chan, J. H. H., Lemon, C., Courbin, F., et al. 2022, A&A, 659, A140
Chandler, C. O., Curtis, A. M., Mommert, M., Sheppard, S. S., & Trujillo, C. A. 2018, PASP, 130, 114502
Chandler, C. O., Kueny, J. K., Trujillo, C. A., Trilling, D. E., & Oldroyd, W. J. 2020, ApJL, 892, L38
Chandler, C. O., Trujillo, C. A., & Hsieh, H. H. 2021, ApJL, 922, L8
Chandra, P. 2018, SSRv, 214, 27
Chen, Z., Zhang, P., Yang, X., & Zheng, Y. 2022, MNRAS, 510, 5916
Cheng, T.-Y., Huertas-Company, M., Conselice, C. J., et al. 2021, MNRAS, 503, 4446
Cheng, T.-Y., Li, N., Conselice, C. J., et al. 2020a, MNRAS, 494, 3750
Cheng, T.-Y., Conselice, C. J., Aragón-Salamanca, A., et al. 2020b, MNRAS, 493, 4209
Chollet, F., & others. 2018, Keras: The Python Deep Learning library, Astrophysics Source Code Library, record ascl:1806.022
Cieza, L. A., Olofsson, J., Harvey, P. M., et al. 2013, ApJ, 762, 100
Cody, A. M., & Hillenbrand, L. A. 2018, AJ, 156, 71
Cody, A. M., Stauffer, J., Baglin, A., et al. 2014, AJ, 147, 82
Conroy, C., Naidu, R. P., Garavito-Camargo, N., et al. 2021, Nature, 592, 534
Conselice, C. J. 2003, ApJS, 147, 1
Corradi, R. L. M., Rodríguez-Flores, E. R., Mampaso, A., et al. 2008, A&A, 480, 409
Cunningham, E. C., Garavito-Camargo, N., Deason, A. J., et al. 2020, ApJ, 898, 4
Curtis, J. L., Agüeros, M. A., Douglas, S. T., & Meibom, S. 2019, ApJ, 879, 49
Dai, J.-M., & Tong, J. 2018, arXiv e-prints, arXiv:1807.05657
Dálya, G., Galgóczi, G., Dobos, L., et al. 2018, MNRAS, 479, 2374

Davenport, J. R. A. 2019, arXiv e-prints, arXiv:1907.04443
Deason, A. J., Belokurov, V., & Evans, N. W. 2011, MNRAS, 416, 2903
Deason, A. J., Belokurov, V., & Koposov, S. E. 2018, ApJ, 852, 118
Denneau, L., Jedicke, R., Grav, T., et al. 2013, Publications of the Astronomical Society of the Pacific, 125, 357
Dey, A., Schlegel, D. J., Lang, D., et al. 2019, AJ, 157, 168
Dey, B., Andrews, B. H., Newman, J. A., et al. 2021, arXiv e-prints, arXiv:2112.03939
Dieleman, S., Willett, K. W., & Dambre, J. 2015, MNRAS, 450, 1441
Dieterich, S. B., Henry, T. J., Jao, W.-C., et al. 2014, AJ, 147, 94
Djorgovski, S. 2000, in Astronomical Society of the Pacific Conference Series, Vol. 213, Bioastronomy 99, ed. G. Lemarchand & K. Meech, 519
Do, A., Tucker, M. A., & Tonry, J. 2018, ApJL, 855, L10
Drew, J. E., Gonzalez-Solares, E., Greimel, R., et al. 2014, MNRAS, 440, 2036
Driver, S. P., Bellstedt, S., Robotham, A. S. G., et al. 2022, MNRAS, 513, 439
Drlica-Wagner, A., Mao, Y.-Y., Adhikari, S., et al. 2019, arXiv e-prints, arXiv:1902.01055
El-Badry, K., Rix, H.-W., & Heintz, T. M. 2021, Monthly Notices of the Royal Astronomical Society, 506, 2269
Engelhardt, T., Jedicke, R., Vereš, P., et al. 2017, AJ, 153, 133
Erkal, D., Deason, A. J., Belokurov, V., et al. 2021, MNRAS, 506, 2677
Fantin, N. J., Côté, P., & McConnachie, A. W. 2020, ApJ, 900, 139
Foreman-Mackey, D., Agol, E., Ambikasaran, S., & Angus, R. 2017, AJ, 154, 220
Forster, F., Cabrera-Vives, G., Castillo-Navarrete, E., et al. 2020, The Astronomical Journal
Freeman, R., & Huang, W. 2014, Nature, 513, 305
Frickel, S., Albert, M., & Prainsack, B. 2016, Investigating Interdisciplinary Collaboration: Theory and Practice Across Disciplines
Gagliano, A., Narayan, G., Engel, A., Carrasco Kind, M., & LSST Dark Energy Science Collaboration. 2021, ApJ, 908, 170
Galli, P. A. B., Bouy, H., Olivares, J., et al. 2021, A&A, 654, A122
Garavito-Camargo, N., Besla, G., Laporte, C. F. P., et al. 2019, ApJ, 884, 51
—. 2021, ApJ, 919, 109
García–Berro, E., & Oswalt, T. D. 2016, New Astronomy Reviews, 72-74, 1
Garnelo, M., D., R., Maddison, C. J., et al. 2018, Conditional Neural Processes. In International Conference on Machine Learning 2018., arXiv:1807.01613 [cs.LG]
Geha, M., Wechsler, R. H., Mao, Y.-Y., et al. 2017, ApJ, 847, 4
Gentile Fusillo, N. P., Manser, C. J., Gänsicke, B. T., et al. 2021, MNRAS, 504, 2707
Gilbert, A. M., & Wiegert, P. A. 2009, Icarus, 201, 714
Giles, D., & Walkowicz, L. 2019, MNRAS, 484, 834
Gizis, J., & Stars, M. W. . L. V. S. C. 2021, LSST Long-Haul Networks (LHN) End-to-end Test Plan
Godines, D., Bachelet, E., Narayan, G., & Street, R. A. 2019, Astronomy and Computing, 28, 100298
Godoy-Rivera, D., Pinsonneault, M. H., & Rebull, L. M. 2021, The Astrophysical Journal Supplement Series, 257, 46
Gómez, F. A., Besla, G., Carpintero, D. D., et al. 2015, ApJ, 802, 128
Gomez, S., Berger, E., Blanchard, P. K., et al. 2020, The Astrophysical Journal, 904, 74
González Hernández, J. I., & Bonifacio, P. 2009, A&A, 497, 497
Gould, A. 2000, ApJ, 542, 785
Goulding, A. D., Greene, J. E., Bezanson, R., et al. 2018, PASJ, 70, S37
Graham, M. J., Drake, A. J., Djorgovski, S. G., Mahabal, A. A., & Donalek, C. 2013, MNRAS, 434, 2629
Graham, M. J., Ross, N. P., Stern, D., et al. 2020, MNRAS, 491, 4925
Graham, M. L., Bosch, J., Guy, L. P., , & the DM System Science Team. 2022, A Roadmap to Photometric Redshifts for the LSST Object Catalog

Grand, R. 2016, in Discs in galaxies (Discs 2016), 34
Greco, J. P., Greene, J. E., Strauss, M. A., et al. 2018, ApJ, 857, 104
Greene, J. E., Strader, J., & Ho, L. C. 2020, ARA&A, 58, 257
Grisel, O., Mueller, A., Lars, et al. 2022, scikit-learn/scikit-learn: scikit-learn 1.1.1
Groot, P., Bloemen, S., & Jonker, P. 2019, in The La Silla Observatory - From the Inauguration to the Future, 33
Guy, J., Astier, P., Baumont, S., et al. 2007, Astronomy & Astrophysics, 466, 11
Hallakoun, N., & Maoz, D. 2021, Monthly Notices of the Royal Astronomical Society, 507, 398
Hamana, T., Shirasaki, M., Miyazaki, S., et al. 2020, PASJ, 72, 16
Hambleton, K., Bianco, F., Clementini, G., et al. 2020, Research Notes of the American Astronomical Society, 4, 40
Han, Y., & Han, Z. 2012, The Astrophysical Journal, 749, 123
—. 2014, The Astrophysical Journal Supplement Series, 215, 2
—. 2018, The Astrophysical Journal Supplement Series, 240, 3
Hartman, Z. D., & Lépine, S. 2020, The Astrophysical Journal Supplement Series, 247, 66
Harvey Wingfield, A., & Myles, R. L. 2014, Sociology Compass, 8, 1206
Hayat, M. A., Stein, G., Harrington, P., Lukić, Z., & Mustafa, M. 2021, The Astrophysical Journal Letters, 911, L33
Heller, R., & Pudritz, R. E. 2016, Astrobiology, 16, 259
Hendel, D., Johnston, K. V., Patra, R. K., & Sen, B. 2019, MNRAS, 486, 3604
Henry, T. J., & McCarthy, Donald W., J. 1993, AJ, 106, 773
Herbig, G. H. 2008, AJ, 135, 637
Hermes, J. J., Kilic, M., Brown, W. R., et al. 2012, ApJL, 757, L21
Hernitschek, N., & Stassun, K. G. 2021, The Astrophysical Journal Supplement Series, 258, 4
Hernitschek, N., Schlafly, E. F., Sesar, B., et al. 2016, The Astrophysical Journal, 817, 73
Herring, C. 2009, American Sociological Review, 74, 208
Heymans, C., Tröster, T., Asgari, M., et al. 2021, A&A, 646, A140
Hill, J. C., Ferraro, S., Battaglia, N., Liu, J., & Spergel, D. N. 2016, PhRvL, 117, 051301
Hillenbrand, L. A., Kiker, T. J., Gee, M., et al. 2022, AJ, 163, 263
Hložek, R., Ponder, K. A., Malz, A. I., et al. 2020, arXiv e-prints, arXiv:2012.12392
Hocking, A., Geach, J. E., Sun, Y., & Davey, N. 2018, MNRAS, 473, 1108
Holman, M. J., Payne, M. J., Blankley, P., Janssen, R., & Kuindersma, S. 2018, The Astronomical Journal, 156, 135
Holoien, T. W. S., Brown, J. S., Stanek, K. Z., et al. 2017, MNRAS, 471, 4966
Hsieh, H. 2015, in IAU General Assembly, Vol. 29, 2251973
Hsieh, H. H. 2009, A&A, 505, 1297
Hsieh, H. H., & Jewitt, D. 2006, Science, 312, 561
Hsieh, H. H., Bannister, M. T., Bolin, B. T., et al. 2019, arXiv e-prints, arXiv:1906.11346
Huber, M., PS1 Science Consortium, & Pan-STARRS IPP Team. 2017, in American Astronomical Society Meeting Abstracts, Vol. 229, American Astronomical Society Meeting Abstracts #229, 237.06
Huertas-Company, M., Gravet, R., Cabrera-Vives, G., et al. 2015, ApJS, 221, 8
Hunter, J. D. 2007, Computing in Science and Engineering, 9, 90
Ishida, E. E. O., Beck, R., González-Gaitán, S., et al. 2019, MNRAS, 483, 2
Ivezić, Ž., Connolly, A. J., VanderPlas, J. T., & Gray, A. 2014, Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data (Princeton University Press)
Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111
Jewitt, D., Hsieh, H., & Agarwal, J. 2015, in Asteroids IV, 221
Jones, D., Scolnic, D., Riess, A., et al. 2018a, The Astrophysical Journal, 857, 51
Jones, R. L., Slater, C. T., Moeyens, J., et al. 2018b, Icarus, 303, 181

Jurić, M., Axelrod, T., Becker, A., et al. 2021, Data Products Definition Document

Jurić, M., Jones, R. L., Kalmbach, J. B., et al. 2019, Enabling Deep All-Sky Searches of Outer Solar System Objects

Kaasalainen, M., & Torppa, J. 2001, Icarus, 153, 24

Kalev, A. 2009, American Journal of Sociology - AMER J SOCIOL, 114, 1591

Kannawadi, A., Hoekstra, H., Miller, L., et al. 2019, A&A, 624, A92

Kanter, R. M. 1977, American Journal of Sociology, 82, 965

Kasen, D. 2010, ApJ, 708, 1025

Kennamer, N., Ishida, E. E. O., Gonzalez-Gaitan, S., et al. 2020, arXiv e-prints, arXiv:2010.05941

Kenworthy, W., Jones, D., Dai, M., et al. 2021, The Astrophysical Journal, 923, 265

Kessler, R., Narayan, G., Avelino, A., et al. 2019, Publications of the Astronomical Society of the Pacific, 131, 094501

Kiman, R., Schmidt, S. J., Angus, R., et al. 2019, AJ, 157, 231

Korol, V., Rossi, E. M., Groot, P. J., et al. 2017, MNRAS, 470, 1894

Kosakowski, A., Kilic, M., Brown, W. R., & Gianninas, A. 2020, ApJ, 894, 53

Koukoufilippas, N., Alonso, D., Bilicki, M., & Peacock, J. A. 2020, MNRAS, 491, 5464

Kovács, G., Zucker, S., & Mazeh, T. 2002, A&A, 391, 369

Kozłowski, S. 2017, A&A, 597, A128

—. 2021, AcA, 71, 103

Kubica, J., Denneau, L., Grav, T., et al. 2007, Icarus, 189, 151

Kunz, M., Bassett, B. A., & Hlozek, R. A. 2007, Phys. Rev. D, 75, 103508

Kusiak, A., Bolliet, B., Ferraro, S., Hill, J. C., & Krolewski, A. 2021, PhRvD, 104, 043518

Lacki, B. C., Brzycki, B., Croft, S., et al. 2021, ApJS, 257, 42

Laine, S., Martinez-Delgado, D., Trujillo, I., et al. 2018, arXiv e-prints, arXiv:1812.04897

Lang, D., Hogg, D. W., & Mykytyn, D. 2016, The Tractor: Probabilistic astronomical source detection and measurement, Astrophysics Source Code Library, record ascl:1604.008

Leahey, E. 2016, Annual Review of Sociology, 42, 81

Leauthaud, A., Singh, S., Luo, Y., et al. 2020, Physics of the Dark Universe, 30, 100719

Léget, P. F., Gangler, E., Mondon, F., et al. 2020, A&A, 636, A46

Lemarchand, G. A. 1994, Ap&SS, 214, 209

Leoni, M., Ishida, E. E. O., Peloton, J., & Möller, A. 2021, arXiv e-prints, arXiv:2111.11438

Lightkurve Collaboration, Cardoso, J. V. d. M., Hedges, C., et al. 2018, Lightkurve: Kepler and TESS time series analysis in Python, Astrophysics Source Code Library, ascl:1812.013

Lindberg, C. W., Huppenkothen, D., Jones, R. L., et al. 2022, AJ, 163, 29

Lochner, M., & Bassett, B. 2021, Astronomy and Computing, 36, 100481

Lochner, M., McEwen, J. D., Peiris, H. V., Lahav, O., & Winter, M. K. 2016, ApJS, 225, 31

Lochner, M., Scolnic, D., Almoubayyed, H., et al. 2022, ApJS, 259, 58

Lomb, N. R. 1976, Ap&SS, 39, 447

Lu, X.-P., & Jewitt, D. 2019, AJ, 158, 220

Lu, Y., Angus, R., Curtis, J. L., David, T. J., & Kiman, R. 2021, VizieR Online Data Catalog, J/AJ/161/189

Lucy, A. B. 2021, PhD thesis, Columbia University, New York

Lucy, A. B., Sokoloski, J. L., Nuñez, N. E., et al. 2018, Research Notes of the American Astronomical Society, 2, 229

Luna, G. J. M., Sokoloski, J. L., Mukai, K., & Nelson, T. 2013, A&A, 559, A6

MacLeod, C. L., Brooks, K., Ivezić, Ž., et al. 2011, ApJ, 728, 26

MacLeod, C. L., Green, P. J., Anderson, S. F., et al. 2019, ApJ, 874, 8

Magrini, L., Corradi, R. L. M., & Munari, U. 2003, in Astronomical Society of the Pacific Conference Series, Vol. 303, Symbiotic Stars Probing Stellar Evolution, ed. R. L. M. Corradi, J. Mikolajewska, & T. J. Mahoney, 539

Malanchev, K. L., Pruzhinskaya, M. V., Korolev, V. S., et al. 2021, MNRAS, 502, 5147

Malz, A., Marshall, P., DeRose, J., et al. 2018, The Astronomical Journal, 156, 35

Mao, Y.-Y., Geha, M., Wechsler, R. H., et al. 2021, ApJ, 907, 85

Margalit, B. 2021, arXiv e-prints, arXiv:2107.04048

Martin, G., Kaviraj, S., Hocking, A., Read, S. C., & Geach, J. E. 2019, Monthly Notices of the Royal Astronomical Society, 491, 1408

Martínez-Galarza, J. R., Bianco, F. B., Crake, D., et al. 2021, MNRAS, 508, 5734

Masci, F. J., Laher, R. R., Rusholme, B., et al. 2019, PASP, 131, 018003

Medan, I., Lépine, S., & Hartman, Z. 2021, AJ, 161, 234

Melchior, P., Moolekamp, F., Jerdee, M., et al. 2018, Astronomy and Computing, 24, 129

Merner, L., & Tyler, J. 2017, AIP Results from 2003–2013 Data of the National Center for Education Statistics

Merton, R., Merton, R., & Company, M. P. 1968, Social Theory and Social Structure, American studies collection (Free Press)

Michalik, D., Lindegren, L., & Hobbs, D. 2015, A&A, 574, A115

Microlensing Software Developers. 2019, Publicly available software for microlensing modeling, simulation and analysis

Millon, M., Courbin, F., Bonvin, V., et al. 2020, A&A, 640, A105

Miret-Roig, N., Bouy, H., Raymond, S. N., et al. 2022, Nature Astronomy, 6, 89

Misra, J. 2020, Research Collaboration: Best Practices.

Moeyens, J., Jurić, M., Ford, J., et al. 2021, AJ, 162, 143

Mróz, P., Street, R. A., Bachelet, E., et al. 2020, Research Notes of the American Astronomical Society, 4, 13

Muir, J., Bernstein, G. M., Huterer, D., et al. 2020, Monthly Notices of the Royal Astronomical Society, 494, 4454

Muthukrishna, D. 2019, in The Extragalactic Explosive Universe: the New Era of Transient Surveys and Data-Driven Discovery, 36

Muthukrishna, D., Narayan, G., Mandel, K. S., Biswas, R., & Hložek, R. 2019, PASP, 131, 118002

Mutlu-Pakdil, B., Sand, D. J., Crnojević, D., et al. 2021, The Astrophysical Journal, 918, 88

Möller, A., & de Boissière, T. 2019, Monthly Notices of the Royal Astronomical Society, 491, 4277

Naidu, R. P., Conroy, C., Bonaca, A., et al. 2020, ApJ, 901, 48

Naiman, J. P., Borkiewicz, K., & Christensen, A. J. 2017, PASP, 129, 058008

Narayan, G., Zaidi, T., Soraisam, M., et al. 2018, Astrophysical Journal Supplement Series

Naul, B., Bloom, J. S., Pérez, F., & van der Walt, S. 2018, Nature Astronomy, 2, 151

Neira, F., Anguita, T., & Vernardos, G. 2020, MNRAS, 495, 544

Newman, J., Blazek, J., Chisari, N. E., et al. 2019, BAAS, 51, 358

Newton, E. R., Mondrik, N., Irwin, J., Winters, J. G., & Charbonneau, D. 2018, AJ, 156, 217

Nidever, D. L., Dey, A., Fasbender, K., et al. 2021, The Astronomical Journal, 161, 192

Nourbakhsh, E., Tyson, J. A., Schmidt, S. J., & The LSST Dark Energy Science Collaboration. 2021, arXiv e-prints, arXiv:2112.07659

Ofek, E. O., Laher, R., Law, N., et al. 2012, PASP, 124, 62

Oguri, M., & Marshall, P. J. 2010, MNRAS, 405, 2579

Olivares, J., Bouy, H., Sarro, L. M., et al. 2021, A&A, 649, A159

O'Mullane, W., & Slater, C. 2020, Schema Management in DM

Pandey, S., Krause, E., DeRose, J., et al. 2021, Dark Energy Survey Year 3 Results: Constraints on cosmological parameters and galaxy bias models from galaxy clustering and galaxy-galaxy lensing using the redMaGiC sample
Park, J. W., Villar, A., Li, Y., et al. 2021, arXiv preprint arXiv:2106.01450
Parker, J. N., & Hackett, E. J. 2012, American Sociological Review, 77, 21
Pasquet, J., Bertin, E., Treyer, M., Arnouts, S., & Fouchez, D. 2019, A&A, 621, A26
Patat, F. 2005, Monthly Notices of the Royal Astronomical Society, 357, 1161
Patat, F., Benetti, S., Cappellaro, E., & Turatto, M. 2006, Monthly Notices of the Royal Astronomical Society, 369, 1949
Pearson, S., Clark, S. E., Demirjian, A. J., et al. 2022, ApJ, 926, 166
Pearson, S., Starkenburg, T. K., Johnston, K. V., et al. 2019, The Astrophysical Journal, 883, 87
Peters, C., Malz, A., & Hlozek, R. 2018, in American Astronomical Society Meeting Abstracts, Vol. 231, American Astronomical Society Meeting Abstracts #231, 245.03
Peters, C. M., & Richards, G. 2015, in IAU General Assembly, Vol. 29, 2257231
Petersen, M. S., & Peñarrubia, J. 2020, MNRAS, 494, L11
Piro, A. L. 2015, ApJL, 808, L51
Piro, A. L., Haynie, A., & Yao, Y. 2021, ApJ, 909, 209
Poleski, R., & Mróz, P. 2018, e-prints
Poleski, R., & Yee, J. C. 2019, Astronomy and Computing, 26, 35
Popinchalk, M., Faherty, J. K., Kiman, R., et al. 2021, The Astrophysical Journal, 916, 77
Portegies Zwart, S., Torres, S., Pelupessy, I., Bédorf, J., & Cai, M. X. 2018, MNRAS, 479, L17
Porter, A. M., & Ivie, R. 2019, AIP Report
Qin, Y.-J., Zabludoff, A., Kisley, M., et al. 2022, ApJS, 259, 13
Qu, H., Sako, M., Möller, A., & Doux, C. 2021, The Astronomical Journal, 162, 67
Raghavan, D., McAlister, H. A., Henry, T. J., et al. 2010, ApJS, 190, 1
Ramírez, I., & Meléndez, J. 2005, ApJ, 626, 465
Rappaport, S., Vanderburg, A., Schwab, J., & Nelson, L. 2021, ApJ, 913, 118
Ren, W., Wang, J., Cai, Z., & Guo, H. 2022, ApJ, 925, 50
Rest, A., Sinnott, B., & Welch, D. 2012, Publications of the Astronomical Society of Australia, 29, 466
Riccio, G., Malek, K., Nanni, A., et al. 2021, A&A, 653, A107
Richards, G. T., Fan, X., Newberg, H. J., et al. 2002, AJ, 123, 2945
Richards, J. W., Starr, D. L., Butler, N. R., et al. 2011, ApJ, 733, 10
Ricker, G. R., Winn, J. N., Vanderspek, R., et al. 2015, Journal of Astronomical Telescopes, Instruments, and Systems, 1, 014003
Rigney, D. 2010, The Matthew Effect: How Advantage Begets Further Advantage (Columbia University Press)
Rix, H.-W., Hogg, D. W., Boubert, D., et al. 2021, AJ, 162, 142
Roberts, E., Lochner, M., Fonseca, J., et al. 2017, JCAP, 2017, 036
Robitaille, T., Beaumont, C., Qian, P., Borkin, M., & Goodman, A. 2017, glueviz v0.13.1: multidimensional data exploration
Rocklin, M. 2015, in Proceedings of the 14th python in science conference No. 130-136, Citeseer
Ross, N. P., Graham, M. J., Calderone, G., et al. 2020, MNRAS, 498, 2339
Rozier, S., Famaey, B., Siebert, A., et al. 2022, arXiv e-prints, arXiv:2201.05589
Sacco, T. 2020, Sociological Forum, 35, 488
Sajadian, S., & Poleski, R. 2019, ApJ, 871, 205
Sánchez-Sáez, P., Lira, H., Martí, L., et al. 2021, AJ, 162, 206
Scargle, J. D. 1982, ApJ, 263, 835
Schaan, E., Ferraro, S., Amodeo, S., et al. 2021, PhRvD, 103, 063513
Schlafly, E. F., Finkbeiner, D. P., Jurić, M., et al. 2012, ApJ, 756, 158
Schmidt, S. J., Wagoner, E. L., Johnson, J. A., et al. 2016, MNRAS, 460, 2611

Schneider, A., Giri, S. K., Amodeo, S., & Refregier, A. 2021, arXiv e-prints, arXiv:2110.02228
Schneider, A., Stoira, N., Refregier, A., et al. 2020, Journal of Cosmology and Astroparticle Physics, 2020, 019
Schuhmann, R. L., Heymans, C., & Zuntz, J. 2019, arXiv e-prints, arXiv:1901.08586
Schwamb, M. E., Jones, R. L., Chesley, S. R., et al. 2018, ArXiv e-prints, arXiv:1802.01783 [astro-ph.EP]
Sergison, D. J., Naylor, T., Littlefair, S. P., Bell, C. P. M., & Williams, C. D. H. 2020, MNRAS, 491, 5035
Sesar, B., Hernitschek, N., Mitrović, S., et al. 2017, The Astronomical Journal, 153, 204
Sheldon, E. S., Becker, M. R., MacCrann, N., & Jarvis, M. 2020, ApJ, 902, 138
Singer, L. P., Parazin, B., Coughlin, M. W., et al. 2022, AJ, 163, 209
Slater, C. T., Harding, P., & Mihos, J. C. 2009, PASP, 121, 1267
Smith, K. L., Mushotzky, R. F., Boyd, P. T., et al. 2018, ApJ, 857, 141
Smith-Doerr, L., Alegria, S. N., & Sacco, T. 2017, Engaging Science, Technology, and Society, 3, 139
Smotherman, H., Connolly, A. J., Kalmbach, J. B., et al. 2021, AJ, 162, 245
Snodgrass, C., Agarwal, J., Combi, M., et al. 2017, A&A Rv, 25, 5
Soga, L., Bolade-Ogunfodun, Y., & Laker, B. 2021, MIT Sloan Management review
Sonnett, S., Kleyna, J., Jedicke, R., & Masiero, J. 2011, Icarus, 215, 534
Sravan, N., Graham, M. J., Fremling, C., & Coughlin, M. W. 2021, arXiv preprint arXiv:2112.05897
Stern, D., Assef, R. J., Benford, D. J., et al. 2012, ApJ, 753, 30
Stern, D., McKernan, B., Graham, M. J., et al. 2018, ApJ, 864, 27
Stetson, P. B. 1996, PASP, 108, 851
Street, R. A., Lund, M. B., Khakpash, S., et al. 2018a, arXiv e-prints, arXiv:1812.03137
Street, R. A., Lund, M. B., Donachie, M., et al. 2018b, arXiv e-prints, arXiv:1812.04445
Tachibana, Y., Graham, M. J., Kawai, N., et al. 2020, ApJ, 903, 54
Tanaka, M., Coupon, J., Hsieh, B.-C., et al. 2018, Publications of the Astronomical Society of Japan, 70, S9
Taylor, G., Lidman, C., Tucker, B. E., et al. 2021, MNRAS, 504, 4111
Taylor, M. B. 2005, in Astronomical Society of the Pacific Conference Series, Vol. 347, Astronomical Data Analysis Software and Systems XIV, ed. P. Shopbell, M. Britton, & R. Ebert, 29
The LSST Dark Energy Science Collaboration, Mandelbaum, R., Eifler, T., et al. 2018, arXiv e-prints, arXiv:1809.01669
The PLAsTiCC team, Allam, Tarek, J., Bahmanyar, A., et al. 2018, arXiv e-prints, arXiv:1810.00001
Thirouin, A., Moskovitz, N., Binzel, R. P., et al. 2016, AJ, 152, 163
Toloba, E., Guhathakurta, P., Romanowsky, A. J., et al. 2016, ApJ, 824, 35
Tröster, T., Mead, A. J., Heymans, C., et al. 2022, A&A, 660, A27
Tsai, C., Corley, E., & Bozeman, B. 2016, Scientometrics, 108, 505, publisher Copyright: © 2016, Akadémiai Kiadó, Budapest, Hungary.
Tsai, K.-H., & Hsu, T. T. 2014, Industrial Marketing Management, 43
Turk, M. J., Smith, B. D., Oishi, J. S., et al. 2011, The Astrophysical Journal Supplement Series, 192, 9
van Roestel, J., Kupfer, T., Bell, K. J., et al. 2021, ApJL, 919, L26
van Velzen, S., Gezari, S., Hammerstein, E., et al. 2021, ApJ, 908, 4
Vanderbosch, Z., Hermes, J. J., Dennihy, E., et al. 2020, ApJ, 897, 171
Vanderbosch, Z. P., Rappaport, S., Guidry, J. A., et al. 2021, ApJ, 917, 41
Vanderburg, A., Rappaport, S. A., Xu, S., et al. 2020, Nature, 585, 363
VanderPlas, J. T. 2018, The Astrophysical Journal Supplement Series, 236, 16
Čvorović-Hajdinjak, I., Kovačević, A. B., Ilić, D., et al. 2022, Astronomische Nachrichten, 343, e210103
Venuti, L., Cody, A. M., Rebull, L. M., et al. 2021, AJ, 162, 101

Venuti, L., Bouvier, J., Flaccomio, E., et al. 2014, A&A, 570, A82
Venuti, L., Bouvier, J., Irwin, J., et al. 2015, A&A, 581, A66
Villar, V. A., Hosseinzadeh, G., Berger, E., et al. 2020, ApJ, 905, 94
Villicaña-Pedraza, I., Carreto-Parra, F., Carramiñana, A., & Saucedo-Morales, J. 2017a, Galaxies, 5, 3
Villicaña-Pedraza, I., Martín, S., Martín-Pintado, J., et al. 2017b, in Formation and Evolution of Galaxy Outskirts, ed. A. Gil de Paz, J. H. Knapen, & J. C. Lee, Vol. 321, 305
Vázquez-Mata, J. A., Hernández-Toledo, H. M., Avila-Reese, V., et al. 2022, Monthly Notices of the Royal Astronomical Society, 512, 2222
Walmsley, M., Lintott, C., Géron, T., et al. 2022, MNRAS, 509, 3966
Ward, C., Gezari, S., Frederick, S., et al. 2021, ApJ, 913, 102
Waszczak, A., Ofek, E. O., Aharonson, O., et al. 2013, MNRAS, 433, 3115
Waszczak, A., Chang, C.-K., Ofek, E. O., et al. 2015, AJ, 150, 75
Waxman, E., & Katz, B. 2017, in Handbook of Supernovae, ed. A. W. Alsabti & P. Murdin, 967
Whidden, P. J., Kalmbach, J. B., Connolly, A. J., et al. 2019, The Astronomical Journal, 157, 119
Willett, K. W., Lintott, C. J., Bamford, S. P., et al. 2013, MNRAS, 435, 2835
Winters, J. G., Henry, T. J., Jao, W.-C., et al. 2019, AJ, 157, 216
Wolf, C., Onken, C. A., Luvaul, L. C., et al. 2018, PASA, 35, e010
Wu, J. F., Peek, J. E. G., Tollerud, E. J., et al. 2022, ApJ, 927, 121
Wyatt, S. D., Tohuvavohu, A., Arcavi, I., et al. 2020, ApJ, 894, 127
Youtie, J., & Bozeman, B. 2014, Scientometrics, 101, 953
Yu, W., & Richards, G. T. 2022, EzTao: Easier CARMA Modeling, Astrophysics Source Code Library, record ascl:2201.001
Zečević, P., Slater, C. T., Jurić, M., et al. 2019, AJ, 158, 37
Zhang, T., Almoubayyed, H., Mandelbaum, R., et al. 2022, arXiv e-prints, arXiv:2205.07892
Zhou, R., Newman, J. A., Dawson, K. S., et al. 2020, Research Notes of the AAS, 4, 181
Zuckerman, H. 1977, Scientific Elite: Nobel Laureates in the United States, Foundations of Higher Education (Free Press)

APPENDIX

A. USE CASE TEMPLATES

We provided science use case and technical use case template for people to start work from at the workshop. These are reproduced here for reference.

A.1. *Science Use Case Template*

Please make a copy of this document before editing (and save as a separate file in Science Use Cases)

Title: <Description of YOUR Science Use Case>
Author: <Name and email>
Date: <Date of whitepaper>

Abstract

Description of your science case. Points to consider:

- Imagine you are writing an abstract for a paper that will be published based on this science use case. What will you discover or measure? Provide a brief background to the science, describe the results that are expected, include any numerical constraints that will come from the analysis (e.g. the distance to which the sources can be detected/analyzed, constraints on cosmological parameters that will be derived etc.)
- What is the size and depth of the dataset that you will use and what type of data will you analyze (e.g. images, catalogs, single band data, light curves etc)
- Comment on why you can't answer your science question today with today's datasets and surveys. What is unique about the LSST data for this science?
- Provide references for background reading for the science and analysis (e.g a couple of references or a pointer to a review article)

Science Objectives

Provide a list of science objectives for the analysis. Points to consider:

- Are there a set of science milestones that are needed to get to the final result? For example, will you need to initially separate stars and galaxies, deblend the photometry, measure photometric redshifts. What are the requirements for each of these individual steps? For example, how well must you measure photometric redshifts to undertake your science case.
- What will limit the science that you can achieve (e.g. the size of the sample, the accuracy of the photometry or astrometry)?
- Is there existing work in this area (e.g. the development of brokers that will analyze and classify the alert streams) that might complement or make use of this use case (e.g. a catalog of supernovae can be used for cosmological distance estimates).

Challenges (what makes it hard)

Describe what are the primary challenges in undertaking the analysis. Points to consider:

- What challenges need to be overcome to undertake this use case. This could be technical challenges (e.g. how to analyze 107 light curves), algorithmic (e.g. current

period finding algorithms are slow and don't work well with poorly sampled data), scientific (e.g. a lack of good models to fit to the data), logistic (e.g. getting access to follow-up telescope time).
- Will the quality of the LSST data impact the analysis (e.g. the number of false positives in the alert stream) and what will be needed to overcome any of these limitations (e.g. writing a specialized real-bogus classifier).
- How often the analysis will be run (e.g. will it be rerun for each Rubin data release or periodically with the alert stream).
- Is there an important timing/urgency constraint on your analysis (e.g. finding candidate microlensing events and triggering space-based followup)? If so, quantitatively, what are those time constraints?

Running on LSST Datasets (for the first 2 years)
What data sets and LSST data products will be analyzed. Points to consider:

- What data sets will you utilize? For example, the alert stream, calibrated images, data release catalogs, the deep drilling fields. How long must the survey be in operation before you will run your analysis (e.g. you need 20 points in a light curve)
- Are the LSST data products sufficient for your analysis or will you need to create value-added catalogs or other derived data products
- What is the size of the data you will use (e.g. the number of light curves you will analyze). Does your science case require analyzing a subset of the population or will you use all galaxies/stars in the data release.

Precursor data sets

- What data can be used today to develop and test these use cases (e.g. the Zwicky Transient Facility (ZTF) public data set). Is this data public?
- Are there other data sets that need to be assembled/collected 'prelaunch' to achieve your science? Do you need a validation set? Will there need to be cross-calibration (e.g. of photometry or astrometry) with those precursor data sets (e.g. gaining longer time baselines from adding LSST to ZTF to Catalina Real-Time Transient Survey (CRTS))?

Analysis Workflow
Provide a step-by-step description of how you will analyze the data including:

- Data cleaning (e.g. removing bad data)
- Derived or intermediate data sets and how these will be stored and accessed
- Matching to existing data sets
- The types of analysis techniques or software packages that will be applied to the data (with a reference)

Software Capabilities Needed

Describe the functionality needed in the software to undertake the science use case. Points to consider:

- Will we need to be able to query the LSST archive and what parameters will be used in the query (e.g. what columns in the database will be used)?
- Are you planning to access your data through community alert brokers? Are there software components (e.g. classification algorithms) needed to enhance these brokers.
- Are there new algorithms that will need to be developed for the science use cases or are there existing software packages that will need modifying for the science use case (e.g. to optimize for speed)?
- Is there new software infrastructure needed to run at the scale of the LSST data and estimates of how quickly these analyses should be run? This could include the ability to access other datasets or cross-match to other catalogs?
- How big are the data sets that will be run on and how slow are current approaches?
- What will need to be stored from these analyses (e.g. for the derived data products) and how much data will that be?
- What functionality will be needed to visualize the data, or the results, or to debug the analysis if it fails?

References for Further Reading

- Provide a short bibliography that provides reference material regarding the science or the analysis.

A.2. *Technical Case Template*

Please make a copy of this document before editing
Title: <Description of YOUR Technical Use Case>
Author: <Name and email>
Date: <Date of this use case>

Abstract
Technical use cases differ from science use cases in that they focus on the development of a specific technique that can be used across multiple applications. Examples of this include photometric redshifts, machine learning tools for finding outliers, tools to predict the orbits of asteroids including uncertainty propagation. For these technical cases the description could include:

- What problems will this technical use case solve? What is the size and depth of the dataset that you will use and what type of data will you analyze (e.g. images, catalogs, single band data, light curves etc).
- Comment on why you can't use existing tools and frameworks. What is unique about the technical development you are proposing?
- Provide references for background reading for the analysis (e.g a couple of references or a pointer to a review article)

Science Cases Needing this Tool
Provide a list of science cases that would utilize this tool or framework.
Requirements for the software
Describe what are the primary requirements and challenges for the software. Points to consider and describe:

- Does the data exist from Rubin in a format that can be used by this software or will new data products need to be generated (see Jurić et al. 2021, Becla 2013 and O'Mullane & Slater 2020). What are the inputs and outputs for the technical analysis (e.g. photometry in multiple bands). Are there specific data structures that would be needed (e.g. light curves)
- What volume of data must be processed and how often
- Is there new software infrastructure needed to run at the scale of the LSST data and what are the estimates of how quickly these analyses should be run? This could include the ability to access other datasets or cross-match to other catalogs?
- Are there other computational challenges that must be addressed (e.g. will memory be an issue when processing the data, will the outputs need to be stored, do database architectures exist for the outputs, will the outputs need to be fed to other packages or to visualization tools).
- What outputs will need to be stored from these analyses (e.g. for the derived data products) and how much data will that be?

- How much temporary storage is needed for processing step outputs which can be deleted on completion?
- Will new functionality be needed to visualize the data, or the results, or to debug the analysis if it fails?
- Why can't we build these tools today (i.e. what makes this use case hard)

Running on LSST and other Datasets
Which data sets and LSST data products will be analyzed. Points to consider:

- What data sets will you utilize? For example, the alert stream, calibrated images, data release catalogs, the deep drilling fields. How long must the survey be in operation before you will run your analysis (e.g. you need 20 points in a light curve)
- Are there precursor data sets on which this analysis can be run and validated. Will analysis of these precursor datasets lead to new publications.

Existing Tools

- What tools exist today to undertake these analyses
- What functionality is missing from these tools and frameworks that would need to be developed, or is there some issue with their application to the dataset at LSST scale? (i.e. what does not work well with these tools, are there missing capabilities or are they too slow or memory-intensive for LSST, etc.).
- If new tools were built what components from existing frameworks would be critical to keep? (e.g. what parts of existing tools work well)

Computational Workflow
Provide a step-by-step description of how you will analyze your data including:

- How will you access the input data? Will we need to be able to query the LSST archive and what parameters will be used in the query (e.g. what columns in the database will be used)?
- Are you planning to access your data through community alert brokers? Are there software components (e.g. classification algorithms) needed to enhance these brokers?
- If you need to use a computational workflow system or resource management (e.g. Pegasus, and Condor) which one would you use and why?
- Describe where existing software packages would be used in each stage of the processing. Are there new algorithms that will need to be developed for the use cases or are there existing software packages that will need modifying for the science use case (e.g. to optimize for speed)?
- Describe whether an analysis will need to be distributed across multiple cores or machines (e.g. can the analysis be undertaken as an embarrassingly parallel application or does it require message passing, is the analysis iterative requiring many passes of the data).

- Do you understand the memory to number of cores ratio needed for the analysis ?
- Describe how the outputs will be stored or visualized
- Describe places where in the workflow there are software tools are missing or won't scale to what you need

References for Further Reading

- Provide a short bibliography that provides a list of the tools or algorithms used in this use case

B. SCIENCE USE CASES

B.1. *Introduction*

During the workshop, participants contributed science use cases that are included here for completeness and as a resource for future work, meant to represent an incomplete sampling of high-priority science for the first two years of LSST. These were organised in major scientific areas; that structure is preserved here as subsections. Within each subsection, one or more use cases are outlined. In this section:

B.2. *Extragalactic static science*

B.2.1. *Low surface brightness dwarf galaxy (candidate) catalog out to 100 Mpc: Multiple Science Cases*

Contributors: Yao-Yuan Mao (yymao.astro@gmail.com), Shany Danieli (sdanieli@astro.princeton.edu)

B.2.1.1. *Abstract*—We will compile a catalog of nearby Low Surface Brightness (LSB) dwarf galaxy candidates using Rubin LSST Year 2 photometric data. These candidates will encompass the majority of dwarf galaxies out to 100 Mpc, down to an r-band apparent magnitude of 24, with >90% purity and >90% completeness. The catalog will include the remeasured photometric properties and distance estimates from a combination of surface brightness fluctuations, spectroscopic follow-up, and photometric redshifts. This census of dwarf galaxies will enable a range of science cases, including mapping of their fundamental distribution functions (e.g., the size–mass plane, the stellar mass function, etc.), and understanding what role the environment plays in shaping these galaxies by comparing isolated ("field") and satellite galaxies. It will also allow stacking of these candidates to enable weak lensing measurements that will map the aggregate dark matter profile of low

mass galaxies, and will provide a list of potential hosts of newly observed supernova and gravitational wave emitters.

B.2.1.2. *Science Objectives* —

- **Science Case 1** - To advance our understanding of dwarf galaxy formation and evolution through a nearly complete sample of dwarf galaxies out to 100 Mpc. In particular, the comparison between field and satellite low-mass dwarf galaxies will allow us to understand how dwarf galaxies evolve, and how quenching mechanisms (e.g., reionization, host galaxy interactions) affect their star formation.
- **Science Case 2** - Dwarf galaxy candidates as lenses: use weak lensing measurement to map the dark matter profile of dwarf galaxies (Sec. 3.2.1 of Drlica-Wagner et al. 2019).
- **Science Case 3** - Host candidates for transient objects.
- **Main Product**: A catalog of dwarf galaxy candidates within 100 Mpc. It should be highly complete (including most true nearby dwarf galaxies) down to a specific r-band apparent magnitude (e.g., $r \leq 24$). It should also have high purity (i.e., contain few galaxies outside the desired redshift limit). For each object in this catalog, we will aim to provide:
  - Photometric properties optimized for LSB dwarf galaxies
  - Distance measurement for a subset, using surface brightness fluctuations (e.g., Carlsten et al. 2019) or spectroscopic follow-up (e.g., Mao et al. 2021)
  - "Photometric distance" for all objects, using an algorithm that is optimized for this distance range (i.e., within 100 Mpc)

B.2.1.3. *Challenges (what makes it hard)* —

- Issues with the source identification and characterization: sky subtraction, deblending/shredding, galactic cirrus.
  - A large fraction of these very nearby dwarf galaxies are very low surface brightness objects, on which the source extraction algorithm may not perform well. At the image level, the sky subtraction and nearby starlight can significantly affect the identification and characterization of these low surface brightness galaxies (e.g., Greco et al. 2018). The main issues include: (1) a galaxy that is not identified by the algorithm at all; (2) a galaxy that is identified as multiple sources (shreds), each of which contains inaccurate photometry; (3) a galaxy that is identified correctly as a single source, but the photometry significantly differs from the true value (e.g., underestimating the luminosity due to missing outskirt light, or overestimating the luminosity due to contamination, or inaccurate sky subtraction); (4) fake sources from galactic cirrus or wrong sky subtraction.
  - Galaxies that are very close (< 5 Mpc) may be partially resolved, that is, some of the stars are identified as point sources, and the rest are identified as diffuse sources (e.g., Mutlu-Pakdil et al. 2021). This case is particularly tricky because

the number of point sources may not be enough for algorithms that search for resolved galaxies (e.g., satellite dwarf galaxies in the Milky Way) to pick them up, and the diffused source may have incorrect photometry.

- Needs an algorithm to estimate the distance of LSB dwarf galaxy candidates using images or improved photometric properties.
  - Existing photometric redshift algorithms do not perform very well in this very low-redshift ($z < 0.05$) regime, if we are aiming for a high completeness and high purity sample. But recent progress has been made on this front (e.g., Wu et al. 2022; Dey et al. 2021).
  - This is difficult mostly due to the lack of training data (dwarf galaxies with known redshift) in the regime that we are aiming for ($z < 0.05$, $r > 20$).
  - To obtain training data, we will need spectroscopic follow-up, which would be expensive. We will need to be strategic about the set of objects that we follow up that can optimize the performance of the algorithm.
  - The mis-characterization in the object catalog (as discussed above) may also mean that the algorithm may need to be run on images (using Convolutional Neural Network (CNN), for example), rather than on catalog entries. Training and running the algorithm on images also requires significant computing resources.
  - We may need a simple algorithm that first selects a high completeness but low purity sample based on catalog-level information only, and another more sophisticated algorithm that runs on images to improve the purity of the sample.

B.2.1.4. *Running on LSST Datasets (for the first 2 years)*—Data sets and LSST data products that will be analyzed:

- Object catalog
- Calibrated images
- Flags from deblending process
- Cross-matched map with a galactic dust map

B.2.1.5. *Precursor data sets*—

- DESI Legacy Imaging Survey, Dark Energy Survey (DES), DECam Local Volume Exploration Survey (DELVE) (these are not as deep, but cover large sky area)
- Hyper Suprime-Cam (HSC)
- Spectroscopic redshift surveys: Galaxy And Mass Assembly (survey) (GAMA) (Driver et al. 2022), DESI, Satellites Around Galactic Analogs (Survery) (SAGA) (Geha et al. 2017; Mao et al. 2021), Merian Survey (Leauthaud et al. 2020)

B.2.1.6. *Analysis Workflow*—

- Visual inspection to understand potential issues with the identification and characterization of the object catalog (hopefully most of this will be done during commissioning)

- Select a high completeness but lower purity sample based on catalog-level information (magnitude, color, size, photometric redshift (photo-z)) only. At this step, we might need to access some deblending flags, so that we can work around issues (shreds, missing sources) in the existing catalog.
- Obtain calibrated cutouts for this sample, re-fit the photometry with models that are optimized for LSB dwarf galaxies. Produce a new catalog with matched photometry and colors.
- Measure distances for a subset of this sample, using surface brightness fluctuations (for sources where this approach is possible), and with spectroscopic cross-match/follow-up.
- Produce "photometric distance" for all objects in the sample, using the improved photometry and images, with the training data from the last step. This "photometric distance" algorithm used here should be optimized for the very low-redshift regime. It may be a repurposed photo-z algorithm, or a ML approach (e.g., CNN) that runs on images directly.
- Improve the purity of this catalog by removing or flagging objects whose distance is greater than 100 Mpc .

B.2.1.7. *Software Capabilities Needed* —

- Image processing software should flag potential issues with sky subtraction, deblending/shredding, and galactic cirrus, so that an add-on program can revisit those regions to find potential LSB galaxies or to re-measure their photometric properties.
- An algorithm that can select a high completeness but low purity sample based on catalog-level information. This output will be used for follow-up distance measurement (e.g., surface brightness fluctuations, spectroscopic redshifts).
- A more sophisticated algorithm that uses all available training data (distance information) to estimate the distance of LSB dwarf galaxy candidates using images or improved photometric properties.

B.2.1.8. *References for Further Reading* —Geha et al. (2017); Greco et al. (2018); Carlsten et al. (2019); Drlica-Wagner et al. (2019); Leauthaud et al. (2020); Dey et al. (2021); Mao et al. (2021); Driver et al. (2022); Wu et al. (2022)

B.2.2. *Stellar streams around external galaxies*

Contributors: Sarah Pearson (spearson@nyu.edu)

B.2.2.1. *Abstract*—With Rubin, we can finally hope to detect stellar streams from accreted dwarf galaxies around external galaxies in a statistical sense. Several other surveys have detected stellar streams around external galaxies, but not to a complete depth nor in a uniform sense. With detection of streams in Rubin LSST, we can measure stream frequency, length, shape, and color, as well as identify potential progenitors. This will allow us to match observations with the full picture of expectations from cosmological hierarchical galaxy formation simulations, thus helping to constrain galaxy formation theories. We will aim to decipher the full accretion histories of galaxies beyond the Milky Way as well as determine the low-mass luminosity function of accreted dwarfs. We will target both massive and dwarf external galaxy hosts, as dwarfs galaxies themselves may have accreted less massive dwarfs. With follow-up radial velocities, Rubin stellar streams may enable us to learn about the dark matter distribution and mass of these external galaxies. Laine et al. (2018) described in detail a project aimed at detecting stellar streams with Rubin LSST, combined with data from other telescopes to help with cirrus and star galaxy separation. This document builds on their science use case and explores what data, software, and analysis will be needed to optimize stellar stream science with Rubin.

B.2.2.2. *Science Objectives*—

- Detect LSB features, such as streams, in Rubin images.
  - Run stream/shell-finding/classification algorithms on LSST data (e.g., Hendel et al. 2019, Pearson et al. 2022)
  - Characterize the lengths, Surface Brightness (SB), widths of detected streams
  - Search for potential progenitor remnants along streams
  - Measure the average color and color profiles of the streams
  - Determine whether there are multiple low surface brightness features (shells/streams) within the host galaxies
- Once we have detections: compare findings to expectations from cosmological simulations of LSB features (e.g., Auriga cosmological simulations, Grand 2016).
- To learn about the potentials of the host galaxies of the streams we might need follow-up radial velocity measurements.

B.2.2.3. *Challenges (what makes it hard)* —

- How do we determine if something is a stream/shell? Detection depends on the contrast with the background. Successful searches will need to comb through huge amounts of data (algorithms for finding such features exist).
- At what points in the 10-year Rubin LSST survey will we be able to detect these features (using estimates of the surface brightness of LSST images + theoretical estimates from simulations)? Estimates of when we should see the brightest vs. when

we should see a complete set of LSB features would be interesting, e.g., comparing detection efficiency in the 1-, 2-, and 10-year stacked image depths.

- We need to handle star galaxy separation in complex image morphology contexts, to avoid contamination of stream photometry by background objects.
- How do we avoid contamination from Galactic cirrus? We may need to restrict targets to those well off the Galactic plane.
- Subtraction of both background objects and the host galaxy is needed to improve the contrast of diffuse structures, and these will have complex image morphologies.
- From Laine et al. (2018): "Reflections and scattered light in the optical path are a real concern for deep wide-field surface photometry. . . . . static star subtraction models applied to stacked images will be insufficient to remove these features. Instead, active-subtraction techniques applied at the data reduction stage (e.g., Slater et al. 2009) must be used to deal with these reflections by modeling and removing them on a star-by-star basis from the individual raw image frames. This makes it imperative that LSST data servers provide users with the raw images, not just the image stacks."
- For multiband images: variation in the number of visits for each band, which will affect the depth reached.

B.2.2.4. *Running on LSST Datasets (for the first 2 years)* —

- While we will detect the brightest streams with just a single visit, more work is needed to characterize the completeness of potential stream detections as a function of stacked depth. We can compare to the expected SB limits of substructure from simulations (e.g., Bullock & Johnston 2005).
  - From Laine et al. (2018): "we have made an approximate estimate of the expected low surface brightness limit that the LSST can provide after the scheduled 825 visits to the same sky location. This corresponds to a total amount of time on source of 3.44h. The expected surface brightness limits will be (3σ; 10x10 arcsec^2 boxes): 29.9 (u), 31.1 (g), 30.6 (r), 30.1 (i), 28.7 (z) AB mag arcsec^{-2} . Each visit that consists of 30 seconds on-source will correspond to the following limits (3σ): 26.6 (u), 27.8 (g), 27.3 (r), 26.8 (i), 25.4 (z) AB mag arcsec^{-2} ."
- Need individual images: "individual images should be made available, not just coadds, as scattered light is much easier to remove from individual frames than from image stacks."
- Ideally, multiband images, maybe other datasets for comparison.
- Multiple postage stamps stacked together, with sizes of postage stamp appropriate for the target galaxy.
- Parallel over stacks of images (process individual galaxy, but all of the images should be available, stacking algorithm is a custom coaddition algorithm (only highest quality data).

B.2.2.5. *Precursor and contemporaneous data sets*—Laine et al. (2018): "Combined LSST/Euclid/WFIRST data set will provide a broad wavelength baseline for the estimation of the ages, metallicities and masses of the stellar populations of disrupted companions"

B.2.2.6. *Analysis Workflow*—

- Remove scattered light from individual images
- Background subtraction to see diffuse structures
- Stacking of multiple postage stamps surrounding galaxies of interest to see connecting features
- Search for streams in data with stream-finding algorithms or by eye (make sure the features are not cirrus, remove false positives)
- Compare to multi-band images or other data sets (e.g, Euclid, SDSS, Roman)
- Characterize lengths, morphology, colors of objects
- Search for progenitor along the detected stream candidates

B.2.2.7. *Software Capabilities Needed*—Need to run stream-finding algorithms (Pearson et al. 2022) on images as well as substructure classification algorithms (Hendel et al. 2019) and determine which are false positives.

B.2.2.8. *References for Further Reading*—Spectroscopic follow-up to Rubin: Newman et al. (2019)
Stellar streams LSST white paper by: Laine et al. (2018)
How to get radial velocities of LSB features: Toloba et al. (2016)
Dwarf companions beyond MW with LSST: https://noirlab.edu/science/sites/default/files/media/archives/documents/scidoc1994.pdf

B.2.3. *Galaxy morphologies for LSST using machine learning with application to photometric redshifts*

Contributors: Ilin Lazar (i.lazar@herts.ac.uk), J. Antonio Vazquez (jvazquez@astro.unam.mx)

B.2.3.1. *Abstract*—Morphology is a fundamental parameter, not only essential for the full spectrum of extra-galactic LSST science but also as a valuable prior in photo-z pipelines that can significantly improve photo-z accuracy (e.g., Pasquet et al. 2019). LSST offers an unparalleled combination of depth, area and statistics, with ≈20 billion galaxies expected from its 18,000 deg^2 footprint with a point-source depth of ≈27.5 mag in the full 10-year stack. This offers a game-changing opportunity to study galaxy morphologies with better precision and statistics than ever before. Since spectroscopy will be sparse, significant investment is being made in photo-z pipelines (e.g., via in-kind contributions by Hatfield/Hsieh et al. in Galaxies and DESC). Improved photo-z measurements (which are essential to LSST science), using morphologies as input, will bring fundamental benefits to the entire LSST community.

A rich literature exists on measuring morphologies in surveys, from visual inspection (e.g., Vázquez-Mata et al. 2022), including those using citizen science systems like Galaxy Zoo (GZ) (Willett et al. 2013), to automated methods, either via simple measures (such as Sersic profile fits, measurement of Concentration/Asymmetry/Smoothness (Conselice 2003), etc.) or sophisticated supervised or unsupervised machine-learning (ML) techniques (e.g., Hocking et al. 2018). However, the unprecedented size of LSST requires a radically different approach. Visual inspection, even using GZ, will be prohibitively time-consuming. Furthermore, since the morphological detail in galaxies will increase as LSST becomes deeper, morphological catalogs will be needed at multiple depths (e.g., from every data release). This makes LSST also challenging for recent supervised ML, since it may be difficult to repeatedly produce large training sets on short timescales, so this has to be well structured in advance. Another solution is unsupervised ML (UML), which can autonomously group morphologically-similar objects into a small number of "morphological clusters", without training sets. Each cluster (rather than millions of individual galaxies) can then be collectively labelled, e.g. into Hubble types, via supervised ML. A member of this project developed such a UML algorithm and validated it on Hyper Suprime-Cam (HSC) surveys (Martin et al. 2019), one of LSST's precursor data sets.

The algorithm randomly samples a large number of patches (of sizes 3-4 pixels squared) in survey images and converts each patch into a 'feature vector' that holds information about its properties (e.g., color/texture). Patches are clustered via a growing neural gas network and hierarchical clustering and galaxies with similar patch properties grouped together. Arbitrarily large galaxy samples, for M > 10^9 $M_\odot$ and $z < 1$ (beyond which classification is difficult from ground-based imaging) are compressed into ∼150 clusters (typical purity >90%). These are easily visually associated with Hubble types and obey known trends in e.g. stellar mass vs star-formation rate. Approximately 150 clusters are needed because identical morphologies at very different redshifts occupy separate clusters (since galaxies

change in size with redshift). Peculiar objects (e.g., mergers) naturally end up in separate clusters. We will provide a classification catalog based on HSC-SSP U/Deep Data Release 3 in 2022 (Lazar+ in prep).

B.2.3.2. *Science/Technical Objectives —*

- Use of ML algorithms to create morphological catalogs for LSST Data Preview 2 (DP2) and subsequent data releases (i.e., DR1, DR2) served through the RSP, using supervised and unsupervised methods.
- Create an RSP interface to enable users to morphologically classify any future LSST data (e.g., varying depth or sky coverage), extending the utility of this project to perpetuity.
- Input the morphologies in the developed codes, demonstrate quality gains and create improved photo-z estimates that can act as a demonstrator for other Rubin photo-z efforts.
- Investigate peculiar/rare objects and the filamentary nature of the Local Universe as a function of galaxy morphology in a statistical sense.

B.2.3.3. *Challenges (what makes it hard) —*

- The algorithm is computationally intensive, and currently takes 10 days to process 1000 square degrees for all objects with $z < 0.5$ on a 4 core/8 GB RAM machine. One could decrease the algorithm processing time and/or computational needs by exploring alternate data management or reduction plans such as using Principal Component Analysis or excluding any data which does not bring any benefit to the clustering procedure. This can be tested using precursor HSC data.
- Propose for CPU time to assure quick classification for every data release. 200 processing cores per data release which amount to 10% of the total LSST computing resources (that can be proposed for) would allow for the algorithm to finish processing in less than 10 days.
- GPU implementation can be used for supervised ML, possibly in an IDAC.
- Need to build the RSP interface in the most robust and user friendly way possible.

B.2.3.4. *Precursor data sets —*

- HSC DR3 DEEP and WIDE (Tested in Lazar in prep.)
- HSC DR2 DEEP (Tested in Martin et al. 2019)
- Hyper Supreme Cam (LSST precursor)
- The HST datasets
- The DESI Legacy Surveys (Dey et al. 2019)
- The LSST Data Previews

These data sets will be beneficial to act as training sets. The algorithm can be run as soon as the first data release is online if unsupervised machine is used, and if training is done on precursor or other external data.

B.2.3.5. *Analysis/Implementation Workflow —*

- We will query galaxy cutouts and catalog properties from the LSST archive. Example catalog properties needed: ra, dec, photometric redshift, stellar mass, SFR, colors, object radius.
- The classification/clustering will be done with algorithms provided by ScikitLearn and suitable processing parallelization will be used.
- The initial training sets will consist of external catalogs and the DP2 images available through the RSP. DP2 will include ~100 deg^2 20-year-depth imaging and ~1600 deg^2 in *g* and *i* bands to Year 1 depth.
- We will test on, and create morphological catalogs for, DP2 and DR1. We will seek guidance about catalog contents from the Galaxies and Informatics and Statistics SCs (via telecons/meetings). The deliverables are DP2 and DR1 catalogs, served through the RSP and described in a tech note.
- Once the first DR comes out, the algorithms (supervised and unsupervised) will be run on these images and morphological catalogs will be generated. About 1Tb of storage will be needed to save these catalogs.
- If 100 cores are used the training timescales may last for a couple of days to a week depending on the depth of the data and sky coverage.
- If GPUs at IDACs are used, the training timescale could be reduced by a factor of 2. However, It will be very important to estimate what is cheaper and more efficient, using more cores within the RSP or moving data to IDACs with GPU facilities.
- RSP interface/documentation – The deliverables are (1) an RSP interface to run our algorithms on any LSST data e.g. by cloning the algorithm's GitHub repository into the LSST structure (or the user's RSP file system) as a library, which the user can import and use via a Jupyter Notebook or Python script, (2) detailed user documentation for using this interface.
- Create catalog for DR2 and forthcoming releases – We will deliver morphological catalogs for different data releases as the LSST survey progresses, served through the RSP.

B.2.3.6. *Software Capabilities Needed —*

- Existing tools:
  - External tools to develop ML algorithms have been developed and optimized to do efficient calculation. Tensorflow, Keras, Pytorch are the most accessible and friendly libraries to work with.
  - The most popular technique used for galaxy classification is convolutional neural networks (CNNs) (e.g., Huertas-Company et al. 2015, Dieleman et al. 2015, Cheng et al. 2021, Walmsley et al. 2022) using training data mainly from the zooniverse. Other techniques are using random forest classifiers, Support Vector Machines (e.g., Goulding et al. 2018) or unsupervised algorithms (e.g., Cheng et al. 2020a). Most of these techniques require large amounts of training data,

are time consuming and may not be able to operate efficiently at LSST scales. Therefore, looking for novel alternatives is necessary.

- What is it needed?
  - An RSP cloud account with Jupyter Notebook and terminal.
  - The algorithm will run using parallel processing.
  - For general use: At least 4 cores, 8Gb of RAM and 10 Gb of storage for any user who wishes to use the ML algorithms on the RSP (for faster processing times 30 cores or more if possible)
  - For large scale catalogue production: Possible accessibility to 200 cores for each LSST yearly release; maybe even have a portion of the LSST CPU capabilities reserved for ML based applications on a yearly basis.
  - Need an efficient architecture within the RSP to be able to import training data in large scales from other surveys.
  - Need for an efficient data transmission to local IDACs with GPU facilities for supervised ML.
- Additional Requirements:
  - Need for combined calibrated images in all bands to carry out the classification.
  - This can be done either by combining images in the RSP during the training process or having previously combined cutouts in PNG format. Meanwhile the first one will require computer power to accelerate the process, the second one will require additional 2 Tb of storage for every 500 millions of 50x50 pix images.

The classification process is expected to be repeated at every data release to generate morphological catalogues.

B.2.3.7. *References for Further Reading* —Cheng et al. (2021) Galaxy finder https://share.streamlit.io/georgestein/galaxy_search); Hayat et al. (2021); Martin et al. (2019); Hocking et al. (2018); Martin et al. (2019); Vázquez-Mata et al. (2022); Willett et al. (2013)

B.2.4. *Estimation of galaxy physical parameters with SED fitting*

Contributors: Gabriele Riccio (facilitator, gabriele.riccio@ncbj.gov.pl), Charlotte Olsen, Raphael Shirley, Sam Schmidt, Julia Gschwend, Viviana Acquaviva

B.2.4.1. *Abstract*—In the past 20 years, the study of the multi-wavelength emission of galaxies from X-rays to radio was found to be necessary to properly analyze the physical properties of galaxies. Because the SED is the result of a complex interplay of several components, such as old and young stars, stellar remnants, the interstellar medium, dust, and supermassive black holes, only the panchromatic view of galaxies can give the full information about their physical properties. To fully comprehend the interactions between these parts, the simultaneous use of different spectral ranges is needed. As broad-band photometry is much less expensive than spectroscopy in terms of observation time, modeling the broad-band SED of galaxies has become one of the most commonly employed methods to evaluate and constrain the physical properties. In this way, properties such as the Star Formation Rate (SFR) and stellar mass, which are essential for a complete understanding of galaxy formation and evolution, can be evaluated. However, modeling the SED can be an intricate problem because galaxies with very different properties can look similar over some wavelength ranges: that is, a young dusty galaxy can appear to be an old dust-free galaxy because they both look red in the optical. This is particularly the case when restricted wavelength ranges, instead of the full SED, are considered. The full SED is rarely available. This makes estimating the physical properties with only a limited wavelength range a great challenge for SED modeling. Considering the large portion of sky that LSST will observe and the depth of forthcoming observations, it is expected that LSST will unveil a significant number of faint galaxies that have remained undetected in current wide-area surveys or that do not have any counterpart in the available multi-wavelengths catalogs. The question is: "how can we use LSST optical observations to obtain estimates of the main physical properties of galaxies, and how realistic and reliable they would be?"

B.2.4.2. *Science Objectives*—

- Fitting of galaxy SEDs
- Estimation of galaxies main physical properties, such as SFR, stellar mass, dust luminosity.
- Test of the reliability of these estimates and flags for possible failures.

B.2.4.3. *Challenges (what makes it hard)* —

- A proper joint estimation of redshift and physical parameters is computationally intensive, and many methods "cheat" and separately estimate a fixed redshift before fitting for the physical parameters. At the very least, that external redshift must be probabilistic rather than deterministic to be meaningful for LSST data, meaning the SED fitting procedures would need to be more advanced than they currently are.
- Many photo-z codes fit with fixed matched aperture photometry, which gives consistent stellar populations within that aperture; however, this may be only a fraction

of the entire galaxy, so fitting for the physical parameters of the total galaxy may be biased by such an aperture approach (this is mainly a problem for very large, extended galaxies with obvious multiple components)
- The limited wavelength coverage of the *ugrizy* filter set limits the robustness of physical parameter estimates compared to those which include Ultraviolet (UV) and infrared (IR) coverage. This will make comparisons with the more extensive data available in the deep drilling fields a key to calibrating SED fitting and physical parameter estimation methods. Extensive documentation of the limitations will also be necessary, given that a large portion of the estimates may be poor for lower S/N objects and those detected in fewer bands.
- Provenance tracking as these fits will be run multiple times (each data release and in between) and we need to store and link the outputs

B.2.4.4. *Running on LSST Datasets (for the first 2 years)*—What data sets and LSST data products will be analyzed. Points to consider:

- Deep drilling fields for validation through multiwavelength counterparts, and general data release catalogs for estimation of the parameters.
- We will be using the Deep Drilling Fields (DDFs) and test on commissioning data. This must be supplemented with IR photometry from legacy surveys in order to constrain galaxy properties (beyond M_* and photo z)
- We will be using a subset of the DDFs such that the galaxies have sufficient Signal to Noise Ratio (SNR) and bands. The size of the dataset will be determined by the number of DDF galaxies for which we have reliable crossmatches in IR datasets

B.2.4.5. *Precursor data sets*—

- As SED fitting requires in general multi-wavelength photometry, up to now several legacy surveys (e.g., Herschel Extragalactic Legacy Project (HELP)) provide mid-far IR photometry of many fields around the sky.
- HSC
- Spectroscopic surveys

B.2.4.6. *Analysis Workflow*—

- Galaxies and stellar identification (to flag in the catalog)
- Preparation of the data for SED fitting (fluxes + photo-z estimates)
- Iterate on quality of the data (visualization in color space to identify outliers)
- Matching to existing data sets where complementary band info exists.
- Undertake the SED fitting with parametric and non-parametric codes (e.g., CIGALE, Prospector).
- Determine quality flags to apply to the data

B.2.4.7. *Software Capabilities Needed*—More sophisticated SED fitters are absolutely needed for LSST to advance our understanding of galaxy populations beyond current capabilities.

Physical parameters like star formation history and stellar mass cannot be separated from redshift inference, or at the very least, inference of the former must account for the nontrivial uncertainty inherent in the latter.

The functionality needed in the software:

- We will query the LSST archive to have measurements (flux, mag, color, size, PSF fit, date), model fit (e.g., point-source, bulge-disk), deblending parameters, aperture surface brightness measurement, photo-z (if available).
- Best-fitting SEDs and tables containing results of the fit need to be stored, resulting in as much data as the LSST data products.
- Custom photometry could be required to adapt LSST images to the same PSF of ancillary data images.

B.2.4.8. *References for Further Reading* —CIGALE: a python Code Investigating GALaxy Emission - Boquien et al. (2019)

Preparing for LSST data. Estimating the physical properties of z < 2.5 main-sequence galaxies - Riccio, G. et al. (2021)

B.3. *Extragalactic transient science*

B.3.1. *Immediate Classification of Astrophysical Transients*

Contributors: Ann Zabludoff (aiz@arizona.edu)

B.3.1.1. *Abstract*—The time-domain community expects millions of new alerts from the Rubin Observatory nightly. After initial filtering, many will be identified as extragalactic, terminal (i.e., explosive) transients. How do we identify which are SNe, Gamma-Ray Burst (GRB) afterglows, TDEs, or more exotic phenomena, before they fade and can no longer be followed up? Tools are necessary to incorporate the properties of the host galaxy, including early LSST imaging, the evolving photometry of the transient, the spatial offset of the transient from its host, the environment of the host, and data from other surveys, to develop rapid probabilistic classifications, which can then be disseminated by LSST event brokers like Arizona-NOIRLab Temporal Analysis and Response to Events System (ANTARES).

B.3.1.2. *Science Objectives*—Here we focus on classifying transients, perhaps even before they occur, solely from prior measurements of their host galaxy properties. Several studies correlating the properties or types of transients with their host galaxies exist (Gagliano et al. 2021; Qin et al. 2022). In general, transients without hosts are cross-matched with galaxy catalogs to find host candidates, and host features are compiled from previous and on-going surveys. Exploration of transient-host connections in these testing and training data will help to optimize large-survey transient brokers by providing transient classifications only from existing space- and ground-measured host galaxy properties. In prioritizing which transients to follow-up with additional observations, such pre-explosion classification is itself valuable, as well as a provider of priors for other classifiers, like those incorporating LSST light-curves, LSST imaging, and data from elsewhere.

Utilizing the known properties of host galaxies is a path to classifying TDEs and other transient types without waiting days, weeks, or months for LSST to provide suitable light-curve data. Yet clear correlations between transient type and host properties have been elusive in past work. For example, while host morphology has been cited as a critical feature, Type Normal Ia SNe occur in both star-forming and dead galaxies, whereas Types II P, Ib, and Ic, which arise from different progenitor stars, are all found in star-forming galaxies. Less common 87A-like and Ic-BL supernovae have been observed in dwarf galaxies, whereas super luminous supernova(e) (SLSN)-II are detected in both dwarfs and more massive hosts. The role of host galaxy metallicity is similarly murky, historically biased by low metallicity dwarfs that dominate the local volume. The relationships of transient types to their host galaxies have not been explored with even a fraction of the data now available nor with a focus on using such links for immediate, i.e., "night of," transient classification.

Several attempts have been made to classify supernovae from host information alone. Gagliano et al. (2021) are able to classify Type Ia and core-collapse supernovae with $\sim 60\%$ accuracy. Gomez et al. (2020) use a specialized algorithm to find SLSNe using host and light curve information, achieving high purity for this rare class. A recent pilot study (Kisley, Ko,

Qin, Zabludoff, & Barnard 2022, ApJ, submitted) using the Qin et al. (2022) database with only a limited feature set and one of many possible Machine Learning (ML) approaches is able to correctly classify 60% of the LSST alerts expected each year for *five* transient classes, including four major types of SNe as well as TDEs. Kisley et al. consider the costs of following up all unclassified transients, finding that one would need to observe only 12-20 of their classified "TDE"s to get one true positive, which is a significant advantage over a random guess, as only <1% of all the alerts are TDEs. The immediate classification provided, even from this initial work, would then allow spectroscopic confirmation and tracking to be achieved early in the time evolution of the TDE, during the super-Eddington phase when the most can be learned about the forming accretion disk and the properties of the supermassive black hole.

Host-galaxy based classifiers need to be updated with additional ML analysis tools, host galaxy features, and rarer classes. Incorporating improved class frequency priors, especially from early LSST data, ancillary data from other surveys, and cooperation with the LSST Brokers are all required to realize LSST's full time-domain science potential.

B.3.1.3. *Challenges (what makes it hard)*—Explosive transient training and testing samples are dominated by the hosts of a few supernova types (i.e., SNe Ia and II account for 85% of the unambiguously classified events). While it is possible to control for this imbalance in determining the significance of the ML results (e.g., Kisley et al.), the deficit of hosts for rarer transient types makes isolating their predictive and distinguishing features difficult. More work is needed to add hosts of long and short-duration GRBs, SLSNe, and subclasses like SNe Ia-02 cx, Ibn, and IIL.

The features of additional uncommon transient hosts can be obtained now with existing facilities. Using 1m-class telescopes, the author of this section and her collaborators have started acquiring SDSS-quality *ugriz* photometry for bright hosts of transient types with typically fewer than 30 known occurrences and that lie just outside the SDSS, Panoramic Survey Telescope and Rapid Response System (Pan-STARRS), and DES footprints. These transients include peculiar SNe Ia (e.g., 1991bg-like and Iax) that are likely produced by exploding white dwarfs, but may not be cosmological standard candles, SNe Ibn that show signs of interaction with a hydrogen-free circumstellar medium, broad-line SNe Ic with ejecta velocities ~3x those of normal SNe, TDEs, and short- and long-duration GRBs. With larger telescopes, one can add missing optical features for fainter galaxies near the SDSS, Pan-STARRS, and DES sensitivity limits that have hosted transient types poorly represented in training and testing samples. We encourage similar follow-up campaigns from the community, in preparation for LSST.

For transient pre-classification in surveys like LSST, Kisley et al. show that it is possible to train a successful ML model solely on archival optical and IR photometric features of known hosts. While such features are the most common in our existing Qin et al. (2022) database, their connections to harder-to-measure, and thus less available, spectroscopic line strengths and derived quantities like stellar mass, star formation rate, and metallicity are ambiguous. Therefore, this work should be expanded to test classification using a hierarchy

of feature subsamples, including spectroscopic and additional photometric features from space mission archives and ground-based galaxy surveys, as well as early LSST data.

We also need to employ a broader range of ML tools, e.g., Real-valued Neural Autoregressive Distribution Estimation (RNADE) or Bayesian Neural Networks, for likelihood estimation. RNADE relies on neural networks to infer the density of features for each transient class (as opposed to kernel density estimation). As a result, it may permit modelling of more complex relationships in the data.

B.3.1.4. *Running on LSST Datasets (for the first 2 years)*—The ML models will improve with training samples that have similar selection to LSST transient-host galaxy pairs, i.e., that use early LSST host-transient data, including rarer transient types. Better understanding of the behavior and location of variable sources like AGN, which early LSST data will provide, is also critical to exclude them when classifying astrophysical transients. After the start of operations, LSST galaxy photometry can be added back into databases like that of Qin et al. (2022) to improve the ML model and prior probabilities for a much more complete sample of potential hosts.

B.3.1.5. *Precursor data sets*—The Qin et al. (2022) transient-host galaxy database covers photometric measurements from the optical to mid-infrared. Spectroscopic measurements and derived physical properties are also available for a smaller subset of hosts. For +36k unique events, the authors have cross-identified ~14k host galaxies using host names, plus +4k using host coordinates. Besides those with known hosts, there are +18k transients with newly identified host candidates. For most hosts, the database contains photometric flux, color, luminosity, surface brightness, concentration, and transient spatial offset. For ~6k events, it also includes host spectroscopic line strengths and derived features such as stellar mass and star formation rate. For ~4k events, it contains X-ray measurements from XMM-Newton, Chandra, Swift, and Röntgensatellit X-ray telescope (ROSAT).

B.3.1.6. *Analysis Workflow*—The transient-host feature database and predictive ML model can be applied to the significant fraction of future LSST target galaxies with existing optical and IR photometric features. For each potential host, one can estimate its probability across the range of transient classes using the training model and prior probabilities based on expected LSST class rates. By cross-matching deep optical catalogs with the UKIDSS and AllWISE catalogs, we estimate that 50% of galaxies brighter than $r \sim 21.34$ have the JHK and $W1$, $W2$ magnitudes required by the Kisley et al. classifiers. There are, on average, nearly 2600 such galaxies per square degree in a typical high Galactic latitude field. Therefore, in the 18000 square degrees surveyed by LSST, there should be about 46 million of these galaxies. Considering the once-per-century supernova rate in typical galaxies, even the Kisley et al. pilot methodology will provide classifications for hundreds of thousands of transients per year. Furthermore, those transients will be generally at lower redshifts where spectroscopic and multi-wavelength follow-ups are more achievable and rewarding. After the start of LSST operations, LSST galaxy photometry can be used to expand the potential host and host feature samples.

B.3.1.7. *Software Capabilities Needed*—Most host galaxy catalogs (e.g., Gagliano et al. 2021; Qin et al. 2022) are rather lightweight, even in FITS format. Rapid/automated access will be critical for follow-up programs for LSST transients due to the enormous number of alerts. ANTARES and Las Cumbres Observatory plan to integrate such standard-format catalogs into their event broker systems.

B.3.2. *In-Depth Studies of Fast Phenomena*

Contributors: Alex Gagliano (gaglian2@illinois.edu)

B.3.2.1. *Abstract*—LSST data will unlock the time-domain sky, with millions of luminous supernovae expected across its decade of operation. A major benefit of scanning the full Southern sky every four nights lies in discovering intrinsically or observationally rare phenomena that have been invisible to us in previous surveys. This includes both rapidly evolving transients for which current samples are small (e.g., Fast blue optical transients (FBOTs), Fast-Evolving Luminous Transients (FELTs)) and rapid early-time phenomena within SN that have been well-characterized at later epochs (e.g., shock breakout, Circum-Stellar Material (CSM) interaction, companion interaction). LSST will observe these signatures, but will only be able to observe an event in a single band every 2–3 weeks. This photometry alone is much too sparse to probe the physical mechanisms underlying these phenomena. To facilitate scientific discoveries in this unique region of parameter space, we must develop software that can identify rapidly-evolving events and prioritize them for high-cadence follow-up with a suite of dedicated resources.

In advance of Rubin first light, effort should be dedicated to exploring the potential pathways to rapid inference of transient events with sparse and noisy multiband LSST photometry. This will start with real-time classification, and explore complementary information that can be used to improve early classification and follow-up prioritization. At present, synthetic light curves from SuperNova ANAlysis (https://snana.uchicago.edu/) (SNANA) within large-scale transient simulations like PLAsTiCC (The PLAsTiCC team et al. 2018) provide the baseline dataset for training classifiers for fast events, but additional work must be done to create software flexible enough to capture realistic and scientifically valuable variation among common events. LSST data will be unique in the diversity of events they will contain (similar in depth and wavelength coverage to the Pan-STARRS Medium Deep Survey; Huber et al. 2017), but significantly larger in scale), and we aim to maximize the scientific yield of these data with complementary software.

B.3.2.2. *Science Objectives*—Accurate LSST photometry at early times is critical for robust inference, as is accurate host galaxy association and photometric redshifts. With this framework in place, LSST data can be used to probe early-time interaction signatures of the supernovae it discovers. These signatures will shed light on a broad range of open questions in supernova science, including the explosion mechanisms driving the majority of SN Ia, signatures of binary companion interaction in SN explosions, the formation of dust in SN ejecta, and the degrees of stripping in stripped-envelope systems. These analyses in the coming years will push us closer to a unified framework for the explosive deaths of stars, and allow us to weave a connecting thread back to the late-stage stellar evolution of their progenitor systems.

B.3.2.3. *Challenges (what makes it hard)* —An ongoing question is the realism of the current generation of transient simulations: classifiers trained with SNANA data typically

underperform on real data for extant surveys, and the reasons for this must be identified and corrected. Alert brokers operating on data from the Zwicky Transient Facility (ZTF) perform some fraction of the tasks described in this science case (follow-up scheduling, real-bogus separation or more sophisticated transient classification), and a good example of this is the ALeRCE postage stamp classifier; but a public end-to-end framework starting from raw photometry to real-time follow-up for fast phenomena remains elusive. Further, the vast majority of existing classifiers are either validated on idealized synthetic datasets or require full-phase light curves for classification and anomaly detection. These softwares are insufficient for finding the fast, early signatures of transient explosion physics and the rapidly decaying events that we hope to study with LSST.

B.3.2.4. *Running on LSST Datasets (for the first 2 years)*—The LSST data necessary for enabling this science are the live alert stream, and if the framework is in place early it could be run at first light (although realistically this will be an iterative process in the first year of the survey stream to improve our pipeline, and this may take the form of active learning to adapt our search in real time). The challenge of this work is conducting this analysis with as few points of LSST photometry as possible – potentially only one or two observations for fast phenomena. Contextual information can aid in classification at early times, and LSST images of the field (which will be available via a postage stamp service within 24 hours of observation) may be valuable for providing these contextual data. To train real-time characterization pipelines in advance of LSST, a synthetic dataset should be constructed consisting of millions of transient events embedded within LSST-like imaging data. This next-generation simulation should encode the transient correlations with their host galaxies that have been previously identified in literature, so that algorithms may investigate the added value of this information.

B.3.2.5. *Precursor data sets*—A variety of existing datasets will be needed to prepare our algorithms for fast transient studies with Rubin. Simulated data from SNANA (transient and host-galaxy photometry) will be useful for testing the value of contextual information for early classification, but synthetic postage stamps would foster additional development and the ZTF alert stream will be critical for validating on real data. The question of enabling follow-up for different science cases requires more than just classification results, and work must also be done in advance to consolidate a large sample of human-labeled "high-priority" events for follow-up to achieve different science goals (e.g., nearby SN II with strange early-time light curve signatures, or an Fast blue optical transient (FBOT) fading quickly). This science driver is made more challenging by the fact that most follow-up decisions made by individual science teams are not made public, and using all available follow-up data may only allow us to do as well as (and not better than) current human-driven follow-up efforts.

B.3.2.6. *Analysis Workflow*—In advance of LSST first light, the general steps to enable this analysis include:

- Generation of a large number of LSST-like light curves with SNANA for both short and long-timescale events. This will include realistic photometric redshifts and observed host-galaxy photometry (globally and at the site of the transient).
- For the longer-timescale events, manipulation of synthetic photometry to include early-time signatures that LSST might observe (informed by theory; see references at the end of this section)
- Fast classification using synthetic properties, with results validated from events discovered within the ZTF alert stream
- Input of raw photometry (event + host) and classification probabilities to an ML method for follow-up prioritization, using prior follow-up decisions to identify "interesting" events even if classification cannot be done at these early epochs.

B.3.2.7. *Software Capabilities Needed*—Multiple pre-existing softwares will need to be combined/repurposed for this task:

- Data will come from alert brokers that have already done initial real-bogus separation and initial downselection of candidates.
- A real-time classifier, probably RNN-based (e.g., RAPID; Muthukrishna et al. 2019) that combines event photometry with contextual information.
- A target selection software that combines raw info (e.g., data directly from brokers) with value-added parameters (classifier results plus priors based on what events observers have found interesting in the past)
- The pipeline should be scalable to millions of alerts and fast (~hours or less for all objects detected per night) in order to facilitate meaningful follow-up.

B.3.2.8. *References for Further Reading*—Some references for the theory of early-time SN signatures include:

- *Shock breakout*: Waxman & Katz (2017)
- *Companion interaction:* Kasen (2010)
- *CSM Shock Cooling:* Piro (2015); Margalit (2021); Piro et al. (2021))
- *Hydrodynamical Interaction of SN Ejecta with Surrounding CSM-Ejecta:* Chandra (2018)

Real-time transient classification:

- Overview of methods: Broccia (2021)
- *RAPID:* Muthukrishna (2019)
- *GHOST:* Gagliano et al. (2021)
- *SuperRAENN:* Villar et al. (2020)
- *Automatic Learning for the Rapid Classification of Events (ALeRCE) Stamp Classifier:* Carrasco-Davis et al. (2021)

B.3.3. *ToO Science (Beyond LIGO GW Triggers)*

Contributors: V. Ashley Villar (vav5084@psu.edu), Igor Andreoni (andreoni@umd.edu), Tomas Ahumada (tahumada@astro.umd.edu), Suvi Gezari (sgezari@stsci.edu)

B.3.3.1. *Abstract*—Little exploration of a target-of-opportunity (Target of Opportunity (ToO)) mode for LSST has been explored beyond LIGO/Virgo fourth observing run (O4) binary neutron star merger followup. Here we enumerate other high-impact ToO observing science cases for LSST and specify the technical challenges which may prohibit these ToO observational modes. We note that for many of these cases, the expected cost (i.e., the impact on other scientific goals) is limited, and the scientific gain may be extraordinarily large.

B.3.3.2. *Science Objectives*—We begin by enumerating the potential ToO triggers enabled by observatories other than the Vera Rubin Observatory:

1. Fast Radio Bursts, the enigmatic new observational discovery of ~millisecond long bursts of radio emission, seemingly from extragalactic origin. It is still unknown if optical emission may accompany these events. LSST is particularly well-matched for FRBs which are non-repeating with low-dispersion measures (i.e., in the nearby universe) and are poorly localized (~degrees uncertainty). Here, a single pointing of LSST (with a few visits) may be sufficient to isolate an optical counterpart. Such localizations are expected regularly in upcoming and ongoing radio observatories such as Canadian Hydrogen Intensity Mapping Experiment (CHIME) and Square Kilometer Array (SKA).
2. Black hole-neutron star (BHNS) mergers. These events are substantially more rare than binary neutron star mergers and often have much larger uncertainty regions. It is additionally not necessarily guaranteed (or even known) if such mergers have an electromagnetic counterpart. BHNS with small localization region (~degrees uncertainty) may be an excellent target for LSST ToOs.
3. Gamma-ray Bursts. Reasonably localized (~degrees uncertainty), nearby gamma-ray bursts would again be well-matched to the field-of-view of LSST for a ToO campaign. Ongoing and upcoming gamma-ray and X-ray observatories, such as Fermi, Swift and Space Variable Objects Monitor (SVOM) should detect dozens of GRBs annually, many of which will likely be well-localized.
4. Finally, we note that neutrino and pulsar timing array hotspots could also be of potential scientific interest.

B.3.3.3. *Challenges (what makes it hard)*—In each of these cases, we must develop ways to cross-match large (degrees) probabilistic areas (over Right Ascension (RA), Dec and potentially redshift) with galaxy/star catalogs.

Simulations for each of these science cases must be developed, and the potential impact of disruptive ToOs should be explored. However, as in the case for Binary Neutron Star

(BNS) ToOs (see references), the impact is likely to be minimal, with the opportunity for highly impactful and timely science.

B.3.3.4. *Running on LSST Datasets (for the first 2 years)* —Within the first two years, template images must be available for efficient search of ToO targets. In a ToO mode, we expect the same alert datastream which will be sufficient for our purposes.

We additionally note that improved galaxy catalogs from LSST (ideally with some photo-z estimate) will greatly our ability to cross-match targets.

B.3.3.5. *Precursor data sets* —ZTF transient alerts. PS1 Medium Deep Survey (PS1-MDS) catalogs and Galaxy List for the Advanced Detector Era (GLADE)(Dálya et al. 2018) galaxy catalogs.

B.3.3.6. *Analysis Workflow* —

- Receive external alert and trigger LSST for a ToO observing mode
  - Highly specialized criteria must be established for automating (ie. democratizing) these triggers.
- Run specialized filters on the ToO alert streams.
  - This can be done via the Brokers, but these specialized filters must be built.
- Cross-match potential candidates with known sources
  - This includes galaxies (especially in the case where redshift probability maps are known) and star catalogs (especially to rule out false positives, such as flaring stars).

B.3.3.7. *Software Capabilities Needed* —A software which can cross-match large probabilistic areas which use HEALPIx.

B.3.3.8. *References for Further Reading* —Singer et al. (2022); Andreoni et al. (2022)

B.3.4. *TDE filtering*

Contributors: Andreja Gomboc (andreja.gomboc@ung.si), Sjoert van Velzen (sjoert@strw.leidenuniv.nl)

B.3.4.1. *Abstract* —TDEs are rare transients, occurring (mostly) in galactic centers. Currently about 10 TDEs are discovered per year. This number will substantially increase with Rubin LSST. Simulations show that we could expect ~10 TDEs detected per night (depending on their rate and LSST observing strategy). Two main challenges are: (1) How to identify a TDE based solely on LSST photometric data? And (2) how to identify a TDE before the peak in the light curve?

We propose a dedicated TDE filter to run on LSST Alert stream data on broker(s). It could be developed in stages: 1) extracting nuclear flares (from the centers of galaxies), 2) photometric feature extraction, 3) photometric typing, including ML. Data it would require would be the information included in the LSST Alerts: history of an object, full photometric light curve, astrometric data (galaxy cross-match, off-set from galactic center), galaxy photo-z, galaxy color/type. Its output would be a stream of nuclear flares with light curve features, including classification labels or probabilities.

LSST with its large Field of View (also denoted field of view (FOV)) (FoV), depth and image quality has the potential to detect many TDEs, enable sample studies, probe Supermassive Black Hole (SMBH) mass distribution, emission processes etc.

B.3.4.2. *Science Objectives* —

- To reliably identify a TDE in order to enable targeted follow-up observations (including spectroscopy), ideally before the peak in the light curve in order to (together with follow-up observations) help determine better the time and magnitude of the peak, and the decay afterwards.
- Infer light curve properties that are helpful in determining the mass of the black hole, type of the disrupted star, impact parameter etc.
- Photometric typing of TDEs using only Rubin data. We want to measure the purity and efficiency of a given filtering approach. The simplest approach is cut on a set of features: rise-time, color, color evolution, and fade timescale. More advanced selection will use ML (either on these features or on the light curve directly).

B.3.4.3. *Challenges (what makes it hard)* —

- Reliable identification of TDEs based solely on Rubin photometric data, which may not have ideal time and multi-band coverage (in particular in u-band filter).
- Measure the purity of our filter using realistic light curves of the background population: SN and AGNs variability.
- Time constraint is to preferably identify a TDE before the peak, i.e. on the order of days to weeks, depending on the time of the first detection.
- Implementation on one or more brokers. Another possibility is to build a TDE Filter for each broker separately.

B.3.4.4. *Running on LSST Datasets (for the first 2 years)* —

- Running on the Rubin Alerts stream. Using available information on galaxies (redshift, type, activity).
- Requirements: reliable separation of stars and galaxies, cross-match with a galaxy, off-set of the transient from the center of a galaxy, photometric redshift, galaxy type.

B.3.4.5. *Precursor data sets* —

- To develop and test a TDE Filter an existing sample of observed TDEs could be used to simulate Rubin TDE alerts with LSST simulation suit.

B.3.4.6. *Analysis Workflow* —

- Data cleaning (e.g. removing bad data)
- Identify nuclear flares (an intermediate data sets)
- Matching to existing data sets (e.g., AGN catalogs)
- Feature extraction from the light curves (e.g., the empirical light curve model in van Velzen et al. 2021)

B.3.4.7. *Software Capabilities Needed* —

- We will run in real-time. The LSST archive might be needed sometimes to update the catalog matching.
- Constructing a nuclear flare sample should be possible on all brokers.
- The light curve feature extraction is more challenging but should also be possible on most brokers. Stress testing will be needed.
- This filtering needs to be tailored for each broker. We therefore plan to pick one broker (Alert Management, Photometry, and Evaluation of Light curves (AMPEL), most likely) and provide the light curve features of the nuclear flares sample as a new data product (either a new is a row-oriented remote procedure call and data serialization framework developed within Apache's Hadoop project (Avro) stream, or via TNS).
- We run on the entire stream. After selection "clear" nuclear flares for the feature extraction, we can expect ~1000 new sources per night. However the active source also have to be updated, so we will quickly be working with 1e5 sources at any given time.
- To visualize the data, we need to be able to pull up the light curve for each nuclear transient including the features. This can be done with Target and Observation Manager (TOM) or Transient Name Server (TNS).

B.3.4.8. *References for Further Reading* —Bricman & Gomboc (2020); and K. Bučar Bricman, S. van Velzen, M. Nicholl, A. Gomboc. "Rubin Observatory's Survey Strategy Performance for Tidal Disruption Events", in prep., ApJ Focus Issue.

B.3.5. *Understand real photometric classification performance*

Contributors: Catarina Alves (catarina.alves.ucl.ac.uk), Fabio Ragosta (fabioragosta@inaf.it)

B.3.5.1. *Abstract*—Traditionally, SN used in astrophysical and cosmological studies need to be spectroscopically classified. However, this will be impossible for most events detected by LSST due to the limited spectroscopic resources; thus, LSST will rely on photometric classification, using the events that will be spectroscopically classified as its training set. Recent efforts, such as the Photometric LSST Astronomical Time-Series Classification Challenge (PLAsTiCC; Kessler et al. 2019) and its follow up ELAsTiCC provide simulated LSST-like datasets of light curves and ancillary data that can be used to train and evaluate photometric classifiers. However, simulations cannot fully model the complexity of real data so the performance of the classifiers can be different on real LSST data. In this analysis, we would use the light curves from the spectroscopically confirmed events of LSST and assess the real performance of the classifiers. Performing this analysis in the first years of LSST allows us to optimize classifiers, including snmachine (Lochner et al. 2016; Alves et al. 2022)), RNN-based classifiers (Muthukrishna et al. 2019; Villar et al. 2020), and CNN-based classifiers (Boone 2019, 2021; Qu et al. 2021) for the remainder of the survey. Moreover, the analysis can give insights if it is needed to further adjust the observing cadence to improve photometric classification of SN.

B.3.5.2. *Science Objectives*—This analysis will provide a framework to compare transient light curve classifiers. It will also lay the ground work to improve said classifiers, e.g., for higher Type Ia purity for cosmological studies. We note that this is especially important for classifiers trained with ZTF data, which is significantly different from the expected LSST data stream.

B.3.5.3. *Challenges (what makes it hard)*—

- Modify the inputs of the classifiers to be able to train with LSST-like data.
- Create a tool to allow users to create custom comparison metrics (e.g., purity of Type Ia SN) using a combination of classifiers.
- Run full light curve analysis can be run periodically (e.g. every year, or each time there is a data release).

B.3.5.4. *Running on LSST Datasets (for the first 2 years)*—The analysis requires light curves from the spectroscopically confirmed events of LSST. It is only relevant after 1 year (after complete, difference imaged light curves are available).

B.3.5.5. *Precursor data sets*—ZTF data can be used as a precursor to evaluate real-data classification. In terms of LSST-like simulated datasets, we have the Photometric LSST Astronomical Time-Series Classification Challenge (PLAsTiCC; Kessler et al. 2019) and its follow up ELAsTiCC.

B.3.5.6. *Analysis Workflow* —

- Generate a method to easily compare/contrast publically available classification algorithms.
- Select classification algorithms and metric to compare algorithms or optimize an ensemble of algorithms.
- Clean the relevant data and extract features accordingly to the chosen algorithm
- Classify the LSST data, and compare classifiers.

B.3.5.7. *Software Capabilities Needed* —

- Alert Brokers will be essential to identify transients in their early phases. Crosstalk between the software and the brokers will enhance the ability of the software to classify the transient. Imperative for the scope – but also to follow up on transients discovered already– will be the ability to reconstruct the history of detection of the source through crossmatch with other catalogs.
- It is also crucial to have access to catalogues of transients in the data releases to perform complete light curve classification and optmise these classifiers.
- There are several existing softwares, which require a common framework to load in data, train, and compare results.
- This work would be applied to archival LSST time series, or ensemble classifiers could be deployed in real time.
- For functionalities related to visualization, see technical case Interactive Data Visualization at scale Section C.8.3

B.3.5.8. *References for Further Reading* —Photometric classifier links: literature review TVS SN from Fabio Ragosta

B.4. *Extragalactic variable science*

B.4.1. *Augmenting AGN Variability*

Contributors: Matthew J. Graham (mjg@caltech.edu)

B.4.1.1. *Abstract* —Variability is one of the quintessential properties of AGN and different enough from that exhibited by other classes of astrophysical object that it can used to identify them in a wavelength agnostic way. The LSST is expected to observe tens of millions of AGN and will thus provide an unprecedented data set for studying population statistics, e.g., tracers of large-scale structure, as well as rare AGN phenomena, such as supermassive black hole binaries, changing-look AGN, and electromagntic counterparts to compact object mergers in AGN accretion disks. However, the timescales of AGN variability typically mean that decadal light curves are required for good characterization and identification and we may be waiting until the 2030s for much of this science.

Bayesian experimental design allows optimizing a set of specified observations to distinguish between different models. In this case, it should be possible to identify sets of additional observations of AGN candidates with external facilities, e.g., ZTF, to maximize the identification of AGN (differentiate between competing statistical models) and reduce the timescale to a viable (good enough) data set.

B.4.1.2. *Science Objectives* —

- Identify a high probability variability-selected AGN data set with minimal temporal coverage from LSST, i.e., as soon as possible.
- Create a complete and uncontaminated catalog to produce an independent estimate of the fraction of LIGO mergers produced by AGN from the excess spatial correlation of AGN and LIGO error volumes.

B.4.1.3. *Challenges (what makes it hard)* —

- The characteristic timescales of AGN variability typically require decadal baseline time series to constrain in an unbiased fashion (Bartos & Kowalski 2017).
- It is unknown what is the minimum amount of data required to differentiate between different statistical/ML models for AGN variability to allow high probability identification.
- It is unknown if multiband data aid in the above or is sparse coverage across six bands a hindrance, i.e., is it better to work in a limited subset of bands?
- It is unknown what are the best models of AGN variability to identify sources as AGN with limited data.
- It is unknown much augmentation from external sources is required to improve LSST data.

B.4.1.4. *Running on LSST Datasets (for the first 2 years)* —

- A subset of AGN can be identified from the alert stream (ZTF statistics suggest about 1% of AGN will show 5 sigma variability).

- The data release catalogs will form the base for the analysis. The analysis will use all galaxies/stars in the data release.
- Deep drilling fields will have a higher cadence but it is not clear whether this helps - specific modeling for this data will be required.
- Other data products, e.g., LIGO probability maps, external catalogs, are essential for this analysis.

B.4.1.5. *Precursor data sets* —

- The Catalina Real-Time Transient Survey and ZTF data can be used to test aspects of this analysis. It is possible that SDSS Stripe 82 data may also help for full multicolor modeling of LSST. All this data is public.
- Preseeding possible AGN candidates from other surveys may bias identification but this also needs to be tested.

B.4.1.6. *Analysis Workflow* —

- Define a base set of AGN variability models to compare, e.g., damped random walk (DRW), damped harmonic oscillator (DHO), Recurrent Neural Network (RNN).
- Define a utility function to compare the success of each model.
- Define a set of possible photometric augmentations.
- Test with the first year of data and second year of data to determin if there is a substantial improvement in AGN identification.

B.4.1.7. *Software Capabilities Needed* —

- This could conceivably run over the full LSST data set but some downscaling to just "variable" sources will be probably happen
- New software infrastructure will be needed to determine best additional observations for subsets of AGN candidates.
- Iterations of fitting AGN models with augmented data will also be needed.

B.4.1.8. *References for Further Reading* —Sravan et al. (2021)

B.4.2. *Conditional Neural Processes for learning AGN light curves*

Contributors: Andjelka B. Kovačević (andjelka@matf.bg.ac.rs), Dragana Ilić (dilic@matf.bg.ac.rs), Paula Sánchez-Sáez (pasanchezsaez@gmail.com), Iva Čvorović Hajdinjak, Robert Nikutta (robert.nikutta@noirlab.edu), Nikola Andrić Mitrović, Mladen Nikolicć (nikolic@matf.bg.ac.rs, Viktor Radović (rviktor@matf.bg.ac.rs), Luka Č Popović (lpopovic@aob.rs)

B.4.2.1. *Abstract*—The next generation time domain surveys, such as Vera Rubin Observatory Legacy Survey of Space and Time (LSST, see Ivezić et al. 2019, and references therein), will provide observations with different cadences over ten years for millions of active galactic nuclei (Bianco et al. 2021; Brandt et al. 2018, AGN) in six filters - *ugrizy*. The consequences of complex, disturbed environments in the vicinity of a supermassive black hole are not well represented by standard statistical models of optical variability in AGN. Thus, developing new methodologies for investigating and modeling AGN light curves is crucial e.g. (Tachibana et al. 2020). Conditional Neural Processes (CNP, Garnelo et al. 2018) are nonlinear function models that forecast stochastic time series based on a finite amount of known data without the use of any object parameters or prior knowledge (kernels). Čvorović-Hajdinjak et al. (2022) provide an initial Conditional Neural Processes (CNP) algorithm that is specifically designed for learning AGN light curves for the intended periodicity pipeline. It was trained using data from:

- the All-Sky Automated Survey for Supernova (ASAS-SN; Holoien et al. 2017), which included 153 AGN
- about 40,000 light curves from Zwicky Transient Facility data release 5 (ZTF DR5; Sánchez-Sáez et al. 2021).

We note that Park et al. (2021) provide a Bayesian version of the CNP, along with an attentive layer. The method simultanously performs regression and reconstruction, with the regression step acting as a regularization on the learned latent space.

Preliminary parallelization experiments show that CNP could efficiently handle large amounts of data. These results imply that CNP could be more effective than standard tools in modeling large volumes of AGN data.In addition to the Multi-Layer Perceptron (MLP) encoder, it is necessary to attempt to develop a One-dimensional (1D) convolution neural process and an attention neural process capable of detecting sequential structure. However, more set stratification testing is required, i.e. light curves in training, validation, and test sets should have a consistent distribution of certain parameters such as time baseline, cadence, gradient, and gaps.

B.4.2.2. *Science Cases Needing this Tool*—

- Estimation of accretion disk and broad line region sizes.
- Estimation of supermassive black hole masses.
- Mining periodicity signal in AGN and stellar light curves.

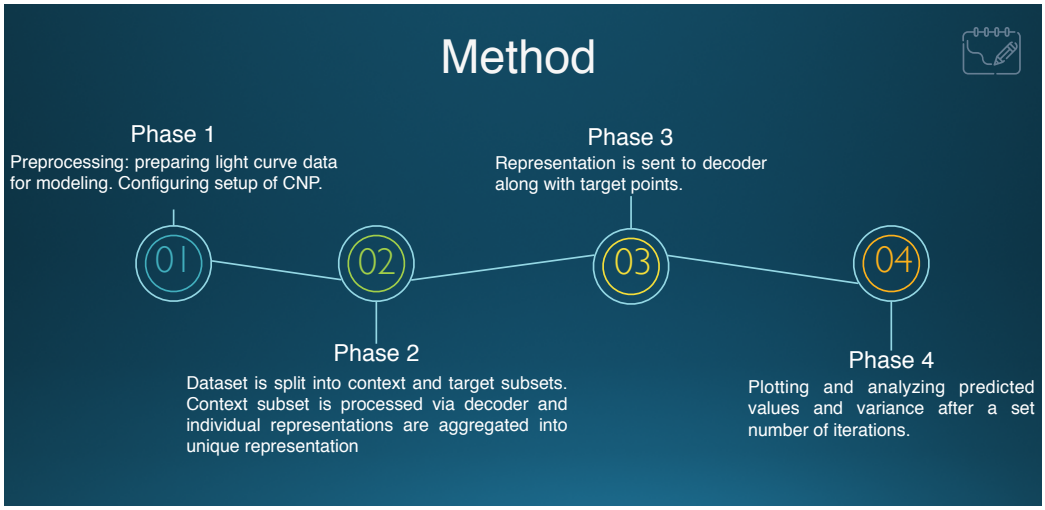

Figure 1. Steps in the conditional neural process method's application. The encoder and aggregator process context points. This process's output is passed to the decoder, together with target points, to calculate predictions. The last unit visualizes both the original and predicted data.

B.4.2.3. *Challenges (what makes it hard)* —

- Light curves show specific phenomena such as flares, possible quasi-periodic oscillations, time gaps, possible irregular cadence due to unpredicted weather conditions,
- Complexity of light curves varies: monotonic light curves, single peaked light curves, more complex (many peaks and valleys).

B.4.2.4. *Running on LSST Datasets (for the first 2 years)* —

- We will apply CNP on the data release light curves. The software will focus on yearly data releases.
- We expect to start modeling light curves in the first few months.
- We will use data from the Deep Drilling Fields as a training set. However, algorithm training will be tested with slower cadences.

B.4.2.5. *Precursor data sets* —CNP has been initially tested on simulated datasets, as well as 153 All-Sky Automated Survey for Supernovae (ASAS-SN) (Holoien et al. 2017) light curves with time baseline about 2,000 days and 40,000 light curves from the Zwicky Transient Facility data release 5 (ZTF DR5) with longer time baseline (Sánchez-Sáez et al. 2021).

B.4.2.6. *Algorithm* —The algorithm runs on data-set O, which consists of n points. Context points , subset O_N with N points, are used for training the neural network, whilst the target points, subset T with M points, are used for making predictions, i.e. modeling the light curve. The input to the encoder is a context set, as illustrated in the Figure 1 (see details in Čvorović-Hajdinjak et al. 2022). The encoder is a multi-layer perceptron (MLP) neural network which produces as output lower dimensional representations (represented with h in Figure 1) $r_i = h((x, y)_i)$ for each context point. The aggregator (a) merges all these representations.

Then these representations are combined with target points x_t (the unlabeled points). Finally, the decoder (a multi-layer perceptron neural network), calculates predictions for each target value and outputs mean and variance of the predicted distribution.

B.4.2.7. *Software Capabilities Needed* —

- Our neural process architecture is developed in Python and uses PyTorch. It consists of two multilayer perceptron neural networks and an aggregator. In addition to MLP encoder, needed to try to implement 1D convolution neural process, and attention neural process that will catch sequential structure.
- To be run on large computing facility based on Graphics Processing Unit (GPU)
- CNP time running complexity is $O(N + M)$ where N is the number of known data points and M is the number of unlabeled data points. It is faster than Gaussian Processes which time running complexity is $O((N + M)^3)$ (Garnelo et al. 2018). However, we found that it is important set stratification, i.e. light curves in training, validation, and test sets should have a comparable distribution of certain parameters such as time baseline, cadence, gradient and gaps. Stratification of training data is in testing phase and demands construction of additional algorithms.

B.4.2.8. *References for Further Reading* —Background on deep learning of AGN light curves: Tachibana et al. (2020)
Background of CNP learning of AGN light curves: Čvorović-Hajdinjak et al. (2022)
Background on Conditional Neural Process: Garnelo et al. (2018)

B.4.3. *Find All the AGN ASAP*

Contributors: Weixiang Yu (editor), K. E. Saavik Ford, Matthew Graham, Yusra AlSayyad, Colin Burke, James Chan, Neven Caplar, Matt O'Dowd

B.4.3.1. *Abstract*—The Rubin Observatory LSST will discover ~100 million AGN after its 10-year survey. Classifying and cataloging those AGN out of a total of ~40 billion LSST sources without the assistance of spectroscopy will be extremely challenging. This challenge will be further elevated by the short baseline (i.e., temporal coverage) at the early times of LSST and a lack of matching multi-wavelength data. Producing a complete and pure sample of AGN as soon as possible (ASAP) once LSST starts is crucial to the timely follow-up of interesting objects (e.g., Changing state quasar or AGN (CSQ)), robust association of Multi Messenger Astronomy (MMA) events, and statistical studies of AGN (e.g., to estimate the luminosity function). Here we present the actual challenges for a complete, pure and timely classification of AGN in LSST and how we may/can overcome them to bring LSST to its full potential.

B.4.3.2. *Science Objectives*—

- Find as many AGN as possible and as soon as possible
- Produce a large sample of highly probable AGN candidates for continuous monitoring and follow-up
- Find as many lensed quasars as possible and at the same time measure the basic lens parameters

B.4.3.3. *Challenges (what makes it hard)*—

- AGN require years-long (even decade-long) baselines to be confidently identified using variability, which LSST will not provide until many years into the survey. Cross-calibrated light curves from precursor surveys (ZTF/DES/Palomar Transient Factory (PTF)/Pan-STARRS/Sloan Digital Sky Survey (SDSS)), for the sake of extending the baseline, are needed to select as many AGN as possible during the early operations of LSST.
- Efficient time-series feature extraction and parametric fitting (e.g., CARMA) for billions of light curves are needed to classify AGN through variability. Light-curve feature extraction should be run at least following each planned data release.
- Optical color selection of Type-1 (unobscured) AGN suffers contamination from stars, where multi-wavelength data can help reduce (stellar) false positives. In addition, successful selection of Type-2 (obscured) AGN relies on multi-wavelength photometry. Efficient large-scale cross-matching between LSST and external multi-wavelength catalogs is a huge technical challenge.
- Deblending of extended AGN (nucleus vs. host galaxy) and lensed AGN.

B.4.3.4. *Running on LSST Datasets (for the first 2 years)*—

- We will utilize both LSST data release catalogs and light curves of mainly all 5-σ sources
- Re-calibrated light curves from precursor surveys and time series features extracted therefrom.
- Cross-matched external catalogs (e.g., eROSITA, Wide-field Survey Explorer (WISE), Gaia, etc.)

B.4.3.5. *Precursor data sets* —

- Time-domain: ZTF/PTF/Pan-STARRS/SDSS/DES/HSC/Gaia
- Multi-wavelength: eROSITA/Visible and Infrared Survey Telescope for Astronomy (VISTA)/WISE/The Very Large Array Sky Survey carried out by Very Large Array (National Radio Astronomy Observatory (NRAO)) (VLA) (VLASS)

B.4.3.6. *Analysis Workflow* —

- Cross-match LSST sources with external catalogs (including time-series and multi-wavelength information)
- Compute time-series features (e.g., parametric fit: CARMA) on joint light curves from precursor surveys and LSST.
- Apply minimum quality cuts.
- Feed LSST catalog data, multi-wavelength photometry, and time-series features into ML algorithms for (probabilistic) classification.

B.4.3.7. *Software Capabilities Needed* —

- Large-scale cross-matching between LSST and external catalogs (including time-series and multi-wavelength information)
- Development of efficient algorithms to compute light curve features (e.g., best-fit CARMA parameters) for billions of light curves.
- Algorithms to select lensed quasars (e.g., CNN on image postage stamps)
- Tools to visualize/interact with selected AGN candidates in a high-dimensional parameter space (e.g., color, variability, proper motion, and others)

B.4.3.8. *References for Further Reading* —Longer baseline justification: Kozłowski (2017, 2021)
More Precursor Surveys: BlackGEM (Groot et al. 2019)
AGN Classification: Richards et al. (2002); MacLeod et al. (2011); Butler & Bloom (2011); Stern et al. (2012); Peters & Richards (2015)

B.4.4. *Connection between short term variability of AGN and their long term behavior*

Contributors: Neven Caplar (ncaplar@princeton.edu), Colin Burke (colinjb2@illinois.edu), K.E. Saavik Ford (sford@amnh.org)

B.4.4.1. *Abstract*—It is well established that individual AGN which are more variable on short timescales (~days to tens of days) are also more variable on long timescales (~years to tens of years). In particular, recent studies have shown deviations from damped random walk at short timescales, with variability being more strongly correlated than expected. It does seem that there is a dependence with mass, with less massive AGN exhibiting a stronger correlation. Are there more connections between short term variability with long term variability, beyond just the amplitude? LSST data promises to deliver AGN curves with relatively high cadence and high precision of measurement. This precision is necessary in order to study relatively small changes of flux at short timescales. A special class of variable AGN are CSQ, i.e., AGN that exhibit large changes of flux and/or change in the spectral properties. The physical mechanism behind these changes is unclear, but high cadence multi-color photometry and 'before and after' spectroscopy are critical to improving our understanding. Predictive models to anticipate state change events would be extremely valuable. We aim to quantify the number of AGN that decrease/increase in luminosity, the number density of sources that change states, as well as their dependence with mass, luminosity and redshift. For changing state AGN it is important to obtain spectra before the change, and follow up spectra (for at least a representative subset of objects). This data will allow us to infer a change in the sound speed and temperature, as well as mass accretion rate as a function of radius from the SMBH. We can also constrain the timescales required for the development and/or collapse of the narrow line region, corona, and disk outflow structures (e.g., warm absorbers).

B.4.4.2. *Science Objectives*—

- Create power spectral density/structure functions for complete list of AGN, previously found in other surveys
- Find the best characterization statistic(s) for AGN variability
- Connect LSST measurements of variability with long term measurements, and find correlations with long term behavior
- In particular, determine short-term variability properties for AGN that exhibit a single large change (CSQ)
- Identify outlier variability behavior – all variability cannot be driven by stochastic gas processes (e.g., microlensing, transits, SN, embedded TDEs, possible MMA sources)
- Identify rapidly-variable AGNs (intermediate-mass black holes as a probe of seeding mechanisms; Dwarf AGN variability for intermediate-mass black hole identification, see Section B.4.6)
- Use CSQ to constrain AGN turn on/turn off process, lifetime of AGN, sound speed/temperature/mass accretion rate as a function of radius, physical drivers of state change, influence of feedback on the rest of a galaxy

B.4.4.3. *Challenges (what makes it hard)*—Challenges may include:

- Large amounts of data (at least 100,000 AGN - if considering only spectroscopically confirmed AGN) - but really want 50 million (and will have to process more to distinguish galaxies on the basis of variability)
- Identifying AGN. This sample will change over time.
- The need for complete datasets (50 million). When we want to estimate Structure Function (SF)[2] or equivalent quantities for the whole dataset, how often do we need to run/re-run this?
- For ML classification, the time and number of light curves is again overwhelming.
- Galaxy contamination, especially for fainter the AGN host galaxy contribution will be significant. This contribution can, in general, be distinguished from variable AGN flux as the galaxy contribution does not vary. But, if we are interested in small changes, small differences in the estimation of galaxy flux (i.e., due to seeing, and chromatic effects) can influence the results. Low luminosity sources cannot be followed up with spectra.
- Low luminosity / dwarf AGN host light contamination for off-nuclear BHs - can we trust those light curves. How do we identify such galaxies with this separation?
- Do we want to capture AGN that exhibit large-changes as they happen, and calculate their short term variability just before the change happened? Is there a way to observe these in almost real time?
- The need to develop predictive models for large changes? This is an ML problem, and whether this is even possible in principle depends on the physical cause of the large changes.
- For small-amplitude, constantly variable sources (AGNs, intermediate mass black holes, off-nuclear AGNs), what are the differential image analysis artifacts?
- For low-luminosity sources, we likely need to determine redshift and mass (and the Eddington ratio) from the variability alone. So we need to figure out how to do that–ML.
- The need to follow-up higher luminosity sources, requires spectroscopic/photometric followup capacity. Some followup might time frames might be as long as 2 or more weeks, others as short as 1 hr (e.g., X-ray observations for quasi-periodic oscillation events).
- Connecting observations to astrophysics requires more theory development and/or ML

B.4.4.4. *Running on LSST Datasets (for the first 2 years)*—

- Primarily will be run on data release catalogs. Several months of data need to be available to create reasonable estimates of short-term variability.
- Value added catalogs would include descriptions of AGN variability on short scales, (e.g., SF^2 estimates at different times; see also the discussion in Bellm 2021)

- The initial dataset could be AGNs that are spectroscopically confirmed (e.g., the SDSS homogeneous sample that is available from the south)
- Large analysis could be done on all objects that exhibit AGN-like variability on time scales of ∼1 yr (i.e., after the first data release). This would lead to a monitoring catalog for e.g., CSQ behavior, and outliers relative to past SF behavior

B.4.4.5. *Precursor data sets* —

- ZTF
- WISE (helps with initial color cuts at the bright end)
- Gaia
- All currently known CSQs
- DES
- Lightcurves generated by Rubin reprocessing of precursor data described in Section B.4.3

B.4.4.6. *Analysis Workflow* —

- Compute variability metric (e.g., total variance) of every object in the data release
- Select the objects that exhibit some minimal amount of variability for one sample
- Select all known AGN for a second sample
- Remove data points that exhibit very large change from previous measurements (in particular if the measurements are in the same night, or only one measurement is very different)
- Select only light curves for which coverage is "reasonably" homogeneous in time domain (e.g., 10 data points, not separated by more than 2 months)
- For each color and/or for all colors jointly, run the power spectral density (PSD) and SF determination algorithm
- Match with known objects, if available, and determine/retrieve long term parameters
- Look for correlations between variability metrics from the modeling done above and long-term parameters

B.4.4.7. *Software Capabilities Needed* —

- Structure function estimation algorithm
- PSD estimation algorithm - e.g., CARMA modeling. Perhaps Gaussian model processing based on code by Dan Foreman-Mackey (EzTao: https://ui.adsabs.harvard.edu/abs/2022ascl.soft01001Y/abstract)? ML modeling algorithm trained on existing data?

B.4.4.8. *References for Further Reading* —Extreme variability Ren et al. (2022)
Stronger dependence of PSD on short time scales Smith et al. (2018)
Timescale-Mass relation Burke et al. (2021a)
Spectroscopic follow up of changing state QSO MacLeod et al. (2019)
Changing state QSO from Catalina Graham et al. (2020)

High redshift changing state QSO Ross et al. (2020)
Changing state QSO in WISE Stern et al. (2018)

B.4.5. *Developing machine learning methods for AGN selection and calculating photometric redshift*

Contributors: Đorđe Savić (djsavic@aob.rs), Isidora Jankov (isidora_jankov@matf.bg.ac.rs), Anđelka Kovačević (andjelka@matf.bg.ac.rs), Dragana Ilić (dilic@matf.bg.ac.rs), Luka Č. Popović (lpopovic@aob.rs), Mladen Nikolić (nikolic.matf@gmail.com), Aleksandra Ćiprijanović (aleksand@fnal.gov)

B.4.5.1. *Abstract* —LSST will produce catalogs for a vast number of sources, which will usher astronomy into a new era of "big data." Machine learning (ML) deployment will be helpful in developing efficient models for various classification and regression tasks. We are focused on three main problems 1) AGN selection, 2) parameterization of AGN light curves and 3) estimating photometric redshifts of AGNs and galaxies. Variability will be the cornerstone for separating AGNs from variable stars. The addition of high quality imaging data will be crucial for separating AGNs from regular galaxies, allowing us to train ML classifiers with superb accuracy > 99%. Redshift estimates for the vast majority of LSST AGNs will rely on photometric estimates. Our goal is to develop an empirical regression method using all the possible sources of information: colors, fluxes, variability, differential chromatic refraction and multiwavelength data.

B.4.5.2. *Science Objectives* —

- AGN selection
- AGN characterization and parametrization of light curves
- Photo-z estimation

B.4.5.3. *Challenges (what makes it hard)* —

- Identifying the variable sources will require iterative processing of time series data for removing artifacts and to improve AGN selection
- Pixel level information will be used for separating galaxies with clear morphological traits from the AGN which requires a lot of computing power when working with convolutional neural networks
- Efficient handling of data quantities greater than that of ongoing surveys
- Running User Defined Functions at scale on LSST data (across all light curves)

B.4.5.4. *Running on LSST Datasets (for the first 2 years)* —

- We will analyze Data Release light curves. Most of the analysis will be based on catalog data
- The initial analysis will process all variable point sources with more than 60 epochs of observations
- We expect to develop a classifying algorithm based on machine learning methods in the first 6 months with further improvements every 6 months.

B.4.5.5. *Precursor data sets* —

- Precursor data are drawn from two main survey fields, an extended SDSS Stripe 82 area and the ESA X-ray Multi-mirror Mission (XMM)-LSS region. The datasets were established by the AGN science collaboration which hosted a data challenge in summer 2021 and are hosted publicly on sciserver (https://www.sciserver.org/). The total amount of objects is of the order of ~450,000.

B.4.5.6. *Analysis Workflow* —

1. Data cleaning and identification of artifacts within the data (this will be iterative as we progressively remove/flag bad data from the time-series catalogs)
    - Remove poorly calibrated photometric data and sources flagged with suspicious photometry (e.g., on edge of CCD or diffraction spikes) and/or LSST DM source quality flags.
    - Remove/flag outlier measurements from a light curve.
2. Data storage and archives
    - The processed data at this stage does not require additional storage.
    - For the photometric redshift estimates, the AGN sources will be matched with the multimessenger data for accurate photometric redshift measurements.
3. Training and applying machine learning methods
    - Select sources with $N > N_{\text{threshold}}$ epochs of data, and within a specific SNR range.
    - Compute the LC features for all bands.
    - Training machine learning methods
    - Classifying all selected sources
    - Further development on machine learning methods used on the accurately classified AGNs for deeper understanding of AGN physics.

B.4.5.7. *Software Capabilities Needed* —

- Ability to apply selection filters to data (Structured Query Language (SQL) query)
- Storage of outputs of filtered and classified data (or flags based on filters applied to existing catalogs)
- Visualization of distributions of selected sources on the sky and relative to camera coordinates
- Visualization of postage stamps and light curves for individual sources
- Visualization of distribution of properties of sources (e.g. color-color scatter plots and histograms) colored by flags

B.4.5.8. *References for Further Reading* —

- Background on machine learning (Berry et al. 2019)
- Deep learning in Python (Chollet & others 2018)

- AGN light curves (Richards et al. 2011)

B.4.6. *Dwarf AGN variability for intermediate-mass black hole identification*

Contributors: Colin J. Burke (colinjb2@illinois.edu)

B.4.6.1. *Abstract* —With a photometric precision of ~percent level or better at g < 22, LSST Rubin has the potential to uncover a population of AGNs in dwarf galaxies ("dwarf AGNs") hosting Intermediate Mass Black Hole (IMBH)s with low luminosities. Owing to their low AGN luminosities and dilution from star-forming host galaxy light, dwarf AGNs have variability amplitudes of a few tenths of a magnitude or less, making them difficult to detect until recently. This proposal is to create a catalog of dwarf AGNs using the first ~year of LSST light curves. Dwarf AGNs with Black Hole (BH) masses smaller than a million solar masses are most variable on days to months timescales, depending on the BH mass. Therefore, the full 10 year LSST light curves are not strictly required.

Recent theoretical and observational results suggest IMBHs may lie outside the nucleus of their host galaxies, so difference images will be an invaluable tool toward identifying these so-called "wandering" off-nuclear AGNs using optical variability. IMBH candidates can be identified from their small-amplitude, stochastic, and short variability timescales using the characteristic variability timescale – BH mass scaling relation. Stellar mass catalogs can be used to select for AGN-like variability in dwarf galaxies.

B.4.6.2. *Science Objectives* —

- Cataloging AGNs in dwarf galaxies and/or AGNs with rapid optical variability (e.g., days), like NGC 4395, to identify IMBH candidates
- Characterize their variability properties and connect with multi-wavelength data
- Enables studies of AGN demographics at low stellar/BH masses which traces early SMBH seeding scenarios

B.4.6.3. *Challenges (what makes it hard)* —

- For small IMBHs with 100 - 10,000 solar masses, the variability power is predicted to be strongest on timescales of ~hours to days (Burke et al. 2021), so getting ~hourly observations for a few days in some fields (e.g., DDFs) would be useful.
- Host galaxy light contamination dilutes variability amplitude
- Inferring the variability timescale in large numbers of light curves is computationally intensive
- Host galaxy stellar mass estimates can be hard if an AGN is involved

B.4.6.4. *Running on LSST Datasets (for the first 2 years)* —

- We will analyze both the alert stream and the data release light curves. Most of the analysis will be based on catalog data (as most low-mass and low-luminosity AGNs are not expected to be variable enough to generate an alert) but we will need to go back to postage stamp cutouts and potentially full fields to visualize any problems with the data.

- The initial analysis will process all variable point sources, identify those with correlated AGN-like variability (e.g., using the Ljung-Box portmanteau test or other test with weak priors on the SF), then identify their characteristic variability timescales (e.g., using the EzTao code; Yu & Richards 2022) or, more likely, piggy-back on AGN variability catalogs and select a sub-sample from them
- Identify opportunities for stellar-mass estimates using multi-wavelength data

B.4.6.5. *Precursor data sets* —

- ZTF (Ward et al. 2021), DES (Burke et al. 2021b), SDSS/PTF (Baldassare et al. 2018, 2020)

B.4.6.6. *Analysis Workflow* —

- Retrieve variability properties (amplitude/timescale)
- Restrict to dwarf galaxies and/or off-nuclear sources in massive galaxies
- Match to external catalogs

B.4.6.7. *Software Capabilities Needed* —

- EzTao (Yu & Richards 2022)

B.4.6.8. *References for Further Reading* —

- IMBH review (Greene et al. 2020)
- Timescale-Mass relation (Burke et al. 2021a)
- Variability-Selection Papers (Baldassare et al. 2018, 2020; Burke et al. 2021b)

B.4.7. *Mapping SMBH Near Fields with Microlensing*

Contributors: Matt O'Dowd (matthew.odowd@lehman.cuny.edu), James Chan (hung-hsu.chan@epfl.ch), Timo Anguita (tanguita@gmail.com), Henry Best (hbest@gradcenter.cuny.edu), Joshua Fagin (jfagin@gradcenter.cuny.edu)

B.4.7.1. *Abstract* —In strongly lensed quasars, the magnification structure in the plane of the source is highly inhomogeneous due to compact bodies in the lensing galaxies, resulting in magnification fluctuations known as microlensing. The rare and brief passage of the quasar central accretion disk across one of these caustic structures results in particularly intense differential magnifications on the scale of the SMBH event horizon. This has the potential to enable tomographic mapping of the structures on that scale. These caustic-crossing events are known to occur but have never been adequately followed up due to their rarity, with average per-lens event frequency <1 per 20 years, and due to the expense of continuous monitoring. High-cadence, multiwavelength monitoring of such an event enable us to measure: 1) the detailed accretion disk structure; 2) the size and geometry of the innermost stable circular orbit and photon sphere, providing a test of general relativity; 3) the geometry and kinematics of the X-ray corona and the X-ray Fe-Kalpha line (also a General Relativity (GR) test).

LSST and the Euclid space telescope are poised to discover >200,000 strong gravitational lenses, including >8,000 lensed quasars. In addition, LSST will monitor the time variability of every lens over its 10-year survey, and has the potential to detect the onset of hundreds of high magnification microlensing events capable of resolving the inner accretion disk each year (~% of the monitored lensed images per annum). The goal would then be to trigger extensive multi-platform follow-up of the most promising events (with, e.g. James Webb Space Telescope (formerly known as NGST) (JWST) + Chandra/XMM - Newton/X-ray Imaging and Spectroscopy Mission (XRISM) + ground-based optical/IR integral field spectroscopy) as the caustic feature scans the inner accretion disk. Due to the volume of the data, the complex behavior of light curves in different filters, and the expense of followup, it is important that we develop a specialized analysis pipeline to confidently identify impending caustic crossing events well in advance of the event itself.

B.4.7.2. *Science Objectives* —

- Studying the inner structure of accretion disk via microlensing light curves
- Predicting the caustic crossing events in lensed quasars through the early LSST data release
- Triggering the extensive follow-ups to capture the caustic crossing events in different filters
- Simulating a comprehensive accretion disk model including microlensing, in order to analyse the observed light curves via machine learning techniques

B.4.7.3. *Challenges (what makes it hard)* —

- Finding the lenses: The first objective is to obtain the largest sample of lensed quasars to date based on LSST imaging, outpacing existing catalogs by at least a factor of about ~20 (Oguri & Marshall 2010). Currently, there are only ~200 lensed quasars found in various surveys. This objective will provide the unprecedentedly large sample size necessary to carry out the subsequent statistical and scientific analyses.
- Timely lightcurve extraction: caustic crossing events can last from ~days to weeks, with a lead-up that can last months. It's important that we identify impending events early (several weeks before the event) to begin intermediate, high-cadence monitoring on smaller telescopes that will be able to confidently define and trigger multi-platform followup timed to the passage of the caustic across the SMBH region.
- Custom deblending: The two or four images of lensed quasars often have sub-arcsecond separation and so will be highly blended. Very careful deblending is critical both for lens discovery and precise photometry, and so we envisage building our own deblending pipeline for all lensed quasars to update lightcurves ~24 hours of observation. Followup triggers will be based on color changes in single lensed images over the first ~few weeks of the microlensing event that leads to a caustic crossing event. Relative photometry at ~a few % between LSST bands is needed to trigger intermediate followup monitoring. An added challenge is that seeing will be different for bands observed on different nights, adding to the importance of careful deblending. Interpolation of lightcurves to enable simulated epoch-matching of photometry in different bands. We note that the flux and relative color changes of an impending caustic-crossing event will not typically generate an alert, necessitating a custom pipeline.
- Follow-up: this program hinges on extensive, multi-platform followup, including by space-based and other high-demand facilities. This will require a suite of companion proposals. It's important that we can demonstrate the reliability of LSST monitoring for the success of these proposals.

B.4.7.4. *Running on LSST Datasets (for the first 2 years)* —

- We'll run our caustic crossing trigger watch on calibrated image cutouts within ~24 hours of observation to ensure timely, accurate deblending
- We'll run our algorithm on LSST coadd images in ugrizy bands to classify the lensed quasar candidates
- Using the LSST imaging, new lenses will be modelled in order to proceed microlensing analysis

B.4.7.5. *Precursor data sets* —

- COSMOGRAIL: database of observed light curves https://www.epfl.ch/labs/lastro/scientific-activities/cosmograil/
- Lensed quasar database: https://research.ast.cam.ac.uk/lensedquasars/index.html
- Gerlumph: containing the LSST light curve and microlensing simulators (https://gerlumph.swin.edu.au)

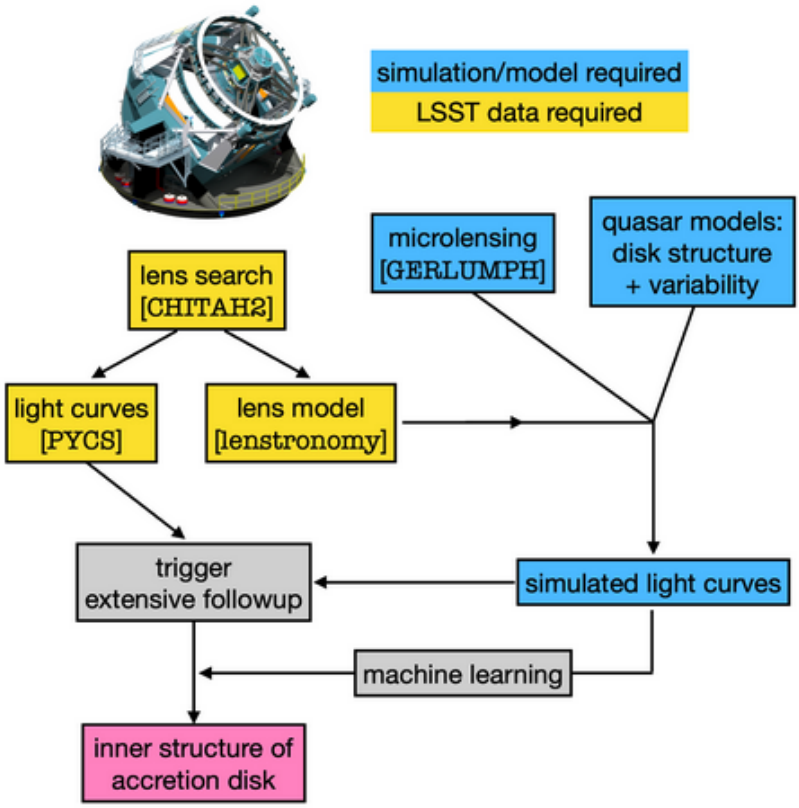

Figure 2. Analysis Workflow - **Figure attribution pending**

- The HSC survey: similar image quality will help build the lens search algorithm (https://hsc.mtk.nao.ac.jp/ssp/survey/)

B.4.7.6. *Analysis Workflow* —See Figure 2.

B.4.7.7. *Software Capabilities Needed* —

- PyCS: extraction of light curves in lensed quasars
- GPU-D: microlensing simulation
- Lenstronomy: lens modelling

B.4.7.8. *References for Further Reading* —

- Lens prediction in LSST (Oguri & Marshall 2010)
- Microlensing light curve simulation in LSST (Neira et al. 2020)
- Light curve extraction (Millon et al. 2020)
- Lens search algorithm (Chan et al. 2022)
- Lens modelling tool (Birrer & Amara 2018)

B.5. *Local universe static science*

B.5.1. *Mapping the Accreted and Intrinsic Stellar Populations in the Milky Way*

Contributors: Nicolás Garavito-Camargo, Adrian Price-Whelan, Alex Riley, Ilija Medan

B.5.1.1. *Abstract*—The Milky Way as a galaxy represents a unique opportunity to constrain the detailed structure and formation history of a galaxy as a way of connecting fine-scale models of galaxy formation to statistical properties of galaxies in the universe. Critical to understanding its formation and evolution is mapping the stellar populations and density structure of the Galaxy as a way of constraining its accretion history, internal structure, and kinematics.

LSST/Rubin will enable mapping the stellar density structure of the Milky Way out to distances never before achieved over large areas of the sky. In particular, the depth of LSST single and co-added images over ~18,000 sq. degrees will provide precise stellar photometric measurements for main sequence stars throughout the outer Galactic disk and inner stellar halo, and will enable tracing evolved stars (Horizontal Branch (HB) and Red Giant Branch (RGB) stars) to the expected edge of the Galactic halo. This precise photometry will then allow measuring photometric stellar parameters (e.g., metallicity and luminosity or distance) for billions of stars, a potential intermediate science product of this work. With stellar population parameters and three-dimensional positional information for these samples, we can (1) map the density structures and asymmetries in the outer Galactic disk as it transitions to the inner stellar halo, (2) model the global structure of the inner halo constrain the clustering scales and stellar populations of known stellar streams and substructures, and (3) chart the density profile and structure of the (largely unconstrained) outer Galactic halo (beyond distances of 150 kpc).

B.5.1.2. *Science Objectives*—

- Catalog tracers - (B)HB stars or otherwise - and their distances (10% accuracy) from the LSST data
- Characterize their distribution around the Milky Way, using both spherically-averaged density profiles and more flexible models
- Connect features in the distribution (e.g., breaks in density profile, global/local under-densities) to the Milky Way's accretion history (e.g., Gaia Sausage-Enceladus (GSE), perturbations from Large Magellanic Cloud (LMC))

With photometry from LSST, we can probe the region that would transition from the outer disk (R > 20 kpc) to the inner halo (R < 100 kpc). To do this, distances and metallicities of the stellar populations must be known to place the populations at this boundary and use chemistry and kinematics to trace the transition zone (Belokurov et al. 2018). When probing this zone we can determine if there is a smooth transition or if these stellar populations are well mixed out to larger radii. Additionally, properties of these populations can be determined as a function of radii to see, for example, if the density structure (i.e. scale height, warp) of the Galactic disk changes at larger radii.

In the above, there are multiple challenges to overcome. To determine photometric metallicities of the sample, a sufficient calibration sample, matched with spectroscopic surveys like 4MOST, GALactic Archaeology with a high-resolution fibre-fed spectrograph for the 1.2m Mercator telescope (HERMES) (GALAH), Large Sky Area Multi-Object Fibre Spectroscopic Telescope, also known as the Guo Shoujing Telescope (LAMOST), etc., must be identified to relate LSST photometry to metallicity of these various stellar populations. Additionally, in the probing of the outer disk we will need to employ crowded-field deblending techniques to ensure a good level of precision (<10%) in the resulting photometric metallicity estimates.

To focus on the inner halo population, in an isotropic universe we expect the inner halo to be defined as a spherically-averaged density profile. Previous studies have shown that this is not true due to the various substructures in the halo (i.e. stellar streams, merger debris, perturbations from dwarf galaxies, etc.). that can cause density enhancement and rapid density variations in the density profile of the halo, such as the break in the halo profile (Deason et al. 2011). It has been shown that the main contribution to the inner halo comes from satellite galaxies that have or are in the process of merging with the MW. In particular, the GSE and Sagittarius (Sag) are the two main accretion events that have contributed to the stellar halo structure (Naidu et al. 2020). LSST will discover many stars in the halo that will belong to these two mergers, such stars will enable the characterization of further details of those mergers, time, mass ratios and orbits. In addition, global properties of the halo, such as shape and mass, can be measured with the shape and kinematics of stellar streams. To identify these substructures, we need to look for over/under-densities (density enhancements) in the inner halo population as compared to a model of a spherically-averaged density profile that has been convolved with the selection function for the survey.

Recently, data from the Gaia satellite along with radial velocities and photometric data have allowed to measure the kinematic structure of the outer stellar halo, proving that the MW is out of equilibrium. The LMC just passed its first pericenter around the MW, during this first approach into the galaxy the LMC is perturbing the galaxy (Garavito-Camargo et al. 2021; Rozier et al. 2022). The phase-space of the inner regions of the MW ($\approx\leq 30kpc$) has been predicted to be displaced by at least 30 km/s and 50 kpc in the last Gyr (Gómez et al. 2015), resulting in an apparent motion of halo tracers with respect to the inner halo. Such 'reflex motion' was measured to be 32 km/s (Petersen & Peñarrubia 2020; Erkal et al. 2021). While the LMC orbits the MW's halo it induces a DM wake trailing the LMC as a result of dynamical friction (Garavito-Camargo et al. 2019), and the counterpart stellar wake has been detected in Conroy et al. (2021). Both the reflex motion and the DM wake are important to characterize in detail since many measured quantities of the MW depend on the dynamical state of the halo (e.g., MW mass, halo triaxiality). LSST's large sky coverage with proper motions and photometry will allow to disentangle the signal from both the reflex motion and the wake with that of substructure in the halo. One of the main challenges of measuring the impact of the LMC in the MW's halo is the presence of substructure that can bias the measured values of the reflex motion (Riley et. al. in prep) and the amplitude

of the wake (Cunningham et al. 2020). LSST's photometry and proper motions will help alleviate these problems by allowing to characterize substructure in the outer halo of the MW.

With the large sky coverage and depth from LSST the structure of density of the outer halo will be revealed. Key science objectives are: 1) Measure density enhancements caused by substructure, perturbations from the LMC, and the smooth halo structure; 2) Determine if the structure of the outer halo similar to that of the inner halo, and if it consists of only substructure and streams or also from mixed structures; 3) See if there is an outer boundary of the outer halo or if the debris fall of continuously. Answering such questions could be addressed with intermediate LSST products such as star catalogs of Horizontal Branch (HB) Stars and RR Lyraes, which have been shown to be a good probe of the density profile of the outer halo (Deason et al. 2011, 2018) . Measuring the extent and density enhancements in the halo will require catalogs that provide 1) star/quasar separation and 2) calibration of distances and metallicities with 10% precision 3) HB stars and RR Lyrae catalogs.

B.5.1.3. *Challenges (what makes it hard)* —Technical challenges: filtering on full catalog of photometry.

Algorithmic:

- Dealing with probabilistic star/galaxy separation
- Dealing with probabilistic stellar parameters
- Constructing appropriate selection functions for density measurements
- Determining detection limits and color dependencies (vs. magnitude)
- Determining efficient selection functions
- Dealing with the effects of extinction by dust
- Dealing with crowding effects
- Making use of only positional data (i.e., how to select halo populations without kinematics or precise chemistry)

Other points:

- How often should the full analysis be run? ~yearly? (gain depth and photometric precision over time)
- No time urgency

B.5.1.4. *Running on LSST Datasets (for the first 2 years)* —

- Create value-added object catalogs with different stellar populations Blue Horizontal Branch (BHB)/RRLyr photometry
- Determine photometric distances for objects in the catalogs
- Cross-match catalogs with previous surveys such as: Gaia, DES, SphereX, SDSS, DELVE, DESI, but sky coverage varies by band
- Determine the selection function
- Analysis could be run every time the new catalogs are published (yearly)

B.5.1.5. *Precursor data sets* —

- Existing mock data don't have substructure / streams
- Could "train" algorithms on Gaia/DELVE catalogs, connect to LSST on bright end, then extrapolate to fainter magnitudes (fainter than g ~20)

B.5.1.6. *Analysis Workflow* —

- Data cleaning
- Star/galaxy separation
- Selection of high-quality stars at the catalog level. Filter using color-magnitude selections
- Matching to existing data sets
- Computing distances using photometry
- Establishing density estimation method that accounts for selection function

B.5.1.7. *Software Capabilities Needed* —

- Ability to query stars from full co-add source catalog and perform color and magnitude selections
- Ability to query selection function / detection probabilities in arbitrarily-refined Hierarchical Equal-Area iso-Latitude Pixelisation (HEALPix) pixelizations of the survey footprint
- Flexible global density modeling with uncertain distances (marginalizations) with 100s of millions of sources
- Traditional integration methods will be slow!
- Flexible mixture density modeling to separate background halo from known structures, like Sagittarius stream
- Need ~few node capability on a compute cluster (handle fairly fast likelihood evaluations on ~100's of millions of sources)
- Ability to install custom software analysis packages on compute resources
- Visualization of large areas of the sky of the stellar halo without extragalactic sources.

B.5.2. *Local Group Dwarf Galaxies Bound and Unbound*

Contributors: Knut Olsen, Adrian Price-Whelan, Alex Riley

B.5.2.1. *Abstract*—The dwarf companion galaxies of the Milky Way, Local Group (LG), and Local Volume (LV) are critical probes both of the dark matter halos that are the seeds of forming galaxies and of the physical processes that shape their formation. This is both because the bound dwarf galaxies directly trace the accretion history and building blocks of the stellar mass in halos, but also because the unbound dwarf galaxies — that now form *stellar streams* that permeate the Milky Way and M31 halos — enable measurements of the detailed dark matter distribution in the Local Group and a full reconstruction of their assembly histories. LSST will provide an extraordinarily rich dataset for the detection and characterization of faint dwarf galaxies and stellar streams, and will enable heavily automated, statistical searches for these classes of objects, allowing for rigorous comparison to theoretical predictions. As described in the report from the LSSTC-sponsored workshop "Searching for Dwarf Companions in the Milky Way and Beyond" (Bechtol 2017), detection techniques for LV dwarf galaxies will include catalog- and pixel-based searches, as well as the use of RR Lyrae as signposts. As described in the report "Probing the Fundamental Nature of Dark Matter with the LSST" from the LSST Dark Matter Group (Drlica-Wagner et al. 2019), the detection of new stellar streams and of density variations in already known streams (a signature of dark matter subhalo interactions) will be possible because of the depth to which LSST can detect main sequence stars around the Milky Way. For this science case, the power of LSST will be its wide, deep, and uniform survey, as well as the complement of the time domain for identifying rare but informative tracers like RR Lyrae stars.

B.5.2.2. *Science Objectives*—

- Identify and characterize the dwarf galaxy population of the Milky Way and nearby galaxies
- Identify and characterize the population of stellar streams around the Milky Way (resolved stars) and nearby galaxies (integrated light and resolved RGB stars)
- Constrain the stellar mass functions and density structures of known dwarf galaxy stellar streams around the Milky Way
- Count the number of streams and measure their properties as a function of radius in the MW
- Characterize and measure the detection probability of dwarf galaxies in catalog- and pixel-based searches for comparison to theoretical predictions
- Connect the population of bound dwarf galaxies and satellites to the population of streams: Which dwarf galaxies and satellites have tidal tails or tidal distortions?

B.5.2.3. *Challenges (what makes it hard)* —

- All current dwarf galaxy and stellar stream searches have humans in the loop, in order to weed out contaminants and false positives (of which there are many)

- Systematic searches and statistical analysis will require automation of detection techniques
- All results will be candidates, requiring spectroscopic information for follow-up. This is expensive and resource-limited
- For measuring stellar density variations in streams, requires knowing the selection function (sum of: detection probability, dustmap, crowding issues, . . .)

B.5.2.4. *Running on LSST Datasets (for the first 2 years)* —

- Catalogs of photometry, ideally at coadd depth
  - Based on sub-selecting metal-poor, distant sources, so ~100's of millions
- Time-domain photometry for a subset of (color-selected) sources (RR Lyrae)
  - ~100,000s of sources
- Coadd images for validation of dwarf galaxy candidates (especially distant ones)
  - ~10,000's of candidates

B.5.2.5. *Precursor data sets* —

- Dark Energy Survey (as an exemplar in semi-automated dwarf galaxy and stream searches)
- Gaia data
- Spectroscopic surveys of streams (S5)

B.5.2.6. *Analysis Workflow* —

- Data cleaning (e.g. removing bad data)
- Derived or intermediate data sets and how these will be stored and accessed
- Matching to existing data sets
- The types of analysis techniques or software packages that will be applied to the data (with a reference)
- Select high-quality stars at the catalog level. Filter using color-magnitude selections
- Run algorithm for finding candidates and quantifying significance
  - Dwarf galaxies - simple/ugali
  - Streams - see software capabilities needed! Existing methods all require humans

B.5.2.7. *Software Capabilities Needed* —

- Real/bogus candidate discrimination for dwarfs (automated)
  - Algorithmic developments
- Automated stream search algorithm
  - Algorithmic developments
  - How to find curved but coherent features that also have distance dependence (on the sky)

  - Would be helpful to have a way of mosaicking large contiguous regions of imaging in an inspectable way, maybe at degraded resolution, for inspection/validation (but that's still human in the loop)

- Ability to query selection function
- Ability to inject artificial sources and re-process (more important for dwarf discovery)

B.5.2.8. *References for Further Reading*—Dwarfs and Milky Way streams in LSST, mainly from a dark matter perspective (Drlica-Wagner et al. 2019)
Dwarf galaxy detections out to 5 Mpc (Mutlu-Pakdil et al. 2021)
Streams in external galaxies with integrated light (Pearson et al. 2019)

B.5.3. *The properties of the faint end of the Main Sequence: the stellar/sub-stellar boundary*

Contributors: Ilija Medan, Knut Olsen, William O'Mullane, Luis Sarro

B.5.3.1. *Abstract*—Low-mass stars are the most abundant stars in the Galaxy (Bochanski et al. 2010) and their long main sequence lifetimes mean they have the potential to trace the entirety of the Galactic star-formation history, and provide clues to understand the structure and evolution of the Milky Way. Additionally, low-mass stars are of great interest for the exoplanet community, as these sources are the best targets for detecting terrestrial planets in habitable zones and knowledge of the chemistry of these stars allows for constraints to be made on the physical properties of the orbiting planets and their formation history. It is also at this faint end where we expect a large increase in non-stellar objects, like white dwarfs and brown dwarfs. For white dwarfs, Fantin et al. (2020) predict 150 million to be observed down to 10-year depth, ~300,000 with 5-sigma parallaxes. So we estimate this dataset to be around 10^9 objects or a significant fraction of the entire object catalog. Here we outline the objectives and challenges in utilizing the LSST photometry to estimate the physical properties of these faint objects, where we specifically discuss the process that may allow for the estimation of properties like mass, radius, metallicity and age. These properties will allow for great progress in many studies that are fundamental to many areas of astrophysics.

B.5.3.2. *Science Objectives*—Physical properties of low-mass stars (mass, radius, metallicity, age) are fundamental to many areas of astrophysics like Galactic archaeology, studies of the IMF and the birth and evolution of star clusters and associations. In this science case we aim to:

- Identify a set of bona fide late-type sources (stellar and substellar) using astrometry and photometry. This is a detection/classification problem where potential contaminants are unresolved galaxies and evolved higher mass stars (giants).
- Estimate basic stellar parameters from the SED: at least the effective temperature ($T_{\rm eff}$). Identify and tag low gravity (young) and low metallicity sources at least in the stellar regime. An example of this kind of analysis for the Calar Alto high-Resolution search for M dwarfs with Exoearths with Near-infrared and optical Echelle Spectrographs (CARMENES) project can be found here.
- Revisit and assess the validity of existing photometric metallicity relations and recalibrate them if needed. Then, infer photometric metallicities for the sample with a recalibration relationship. This is only possible for the stellar regime.
  - A good starting point is, e.g., Ramírez & Meléndez (2005), Schmidt et al. (2016), Medan et al. (2021).
- A recalibration will require the identification of members in binary systems with bright primaries of well determined metallicities due to difficulties in determining metallicities of low-mass stars from spectra (see Allard et al. 1997).

- Characterize the selection function so that inference at the level of population distributions can be achieved. This is crucial for studies of the luminosity and mass functions and their dependence on age and location in the local Galaxy.
  - Potential selection filters appear in the limiting magnitude (survey design) and the classification stage described above.
- The derivation of luminosities, masses and radii using several sets of theoretical models and revisit the possible existence of a radius minimum beyond the stellar/sub-stellar boundary where degeneracy pressure takes over as supporting mechanism (e.g. Dieterich et al. 2014).
- Recalibrate mass-luminosity and radius-luminosity relations.
- Evaluate, where possible, age estimation techniques like gyrochronology (requires time series modeling to infer rotation periods; see e.g. Popinchalk et al. 2021; Godoy-Rivera et al. 2021). Evaluate alternative methods like hierarchical models that infer age and mass distributions as hyperpriors.
- Identifying chemo-kinematic sub-populations based on the astrometric properties (proper motions and parallaxes) and the metallicities estimated above. See, for example, Hallakoun & Maoz (2021).
- Identify new faint companions to existing binaries (e.g., from Hartman & Lépine 2020; El-Badry et al. 2021) and identify new binary systems using astrometric methods. With these detections, can study the multiplicity rate of stellar systems (compare to e.g. Raghavan et al. 2010; Winters et al. 2019) and of sub-stellar systems.

See references for further reading to see prior work.

B.5.3.3. *Challenges (what makes it hard)* —There are multiple challenges with accomplishing the above goals. These include:

- Understanding selection effects, which is an issue and a general problem for many science cases. Particularly, we must understand the effect selecting these stars from the full set and what the level of either misclassification or over-classification (i.e., being too restrictive in our cuts) has on the overall sample.
- Generally, understanding how this work will interact with dynamic stellar science.
- Getting a large enough calibration subset for a large range in mass. There can be a challenge in getting a large enough sample that truly covers the mass range of M dwarfs, leading to not well constrained results at the lower mass end.
- Removing contaminants (i.e., unresolved binaries and active stars) from the calibration subset. This is most easily done with parallax measurements (to spot clearly over luminous sources that would have poorly estimated physical parameters from spectra). While this can be done if the pair is in Gaia (which may be a requirement for identifying wide binaries used to calibrate metallicity), this means the same kind of cleaning may not be able to be done once the photometric metallicity relationship starts to be used on field stars that may not have parallax measurements.

- Photometric metallicity calibrations may not be as accurate for M dwarfs when just using optical data (see Schmidt et al. 2016; Medan et al. 2021).
- Large number of light curves to extract properties from 10^9 objects.
- Need more epochs to constrain this problem (i.e. identify contaminants and get astrometry), so need more years of data. Additionally, overall quality of data will improve later in the mission as we probe to fainter and fainter magnitude limits.
- Star-Galaxy distinction for these low mass stars will be an issue.

B.5.3.4. *Running on LSST Datasets (for the first 2 years)*—For this work, we will mainly use the object catalog data to estimate physical parameters and will do so on every data release. For detecting variable stars and estimating ages using rotational properties, we will also use the source data to extra data via light curve analyses. We do note that LSST data products will not be entirely sufficient for the first two years and we will need to extrapolate further properties (like proper motion) with external catalogs for the initial releases.

B.5.3.5. *Precursor data sets*—We will utilize multiple precursor datasets throughout the analysis. For astrometry during the first two years, we will utilize data from Gaia and Dark Energy Camera (DECam) (which will be used to bootstrap proper motions). Additionally we will need data from large spectroscopic surveys (e.g., SDSS, GALAH) for the metallicity calibration samples.

B.5.3.6. *Analysis Workflow*—In the analysis for this project, we can generally split the methodology for how properties are estimated into those that only need color and those that need color and astrometry. We expect that we can classify objects with color alone and also maybe metallicity, though it will not be as precise. Then, we expect that we will also need astrometry to determine multiplicity, luminosity, mass and radius, and a more precise estimate of metallicity.

For the above, we expect the steps in the analysis to be the following for all properties:

1. Make cuts on the object catalog to classify faint objects. Initial cuts will be made with color alone, and more precise classification with early proper motion and parallax measurements.
2. Get the source data for all objects identified in the object catalog. Use light curves built with the source data to:
    - Identify variable stars
    - Identify prime candidates for age calibration
3. Cross-match the above samples with external catalogs
4. Construct a calibration subset for each physical property of interest. The calibration subset should evenly span the mass range probed.
5. Calibrate relationships for each physical property. These calibrations will vary for each physical property.
6. Apply the calibrated relationships to the objects initially selected from the object catalog.

7. Repeat and refine these calibrations with subsequent data releases.

B.5.3.7. *Software Capabilities Needed*—For the proposed science case, the main software capabilities needed are:

- The ability to query the LSST archive based on a combination of color, proper motion and parallax.
- The ability to quickly process 10^9 light curves.
- Have the ability to cross-match the data to other large astronomical surveys.
- Create new tables of physical properties for the sample.

We believe that most of the above can be done with existing infrastructure as the above queries can be written in SQL and executed with Qserv. The exception to this may be with the processing of the large number of light curves for this sample.

B.5.3.8. *References for Further Reading*—Bellm (2021): covers Rubin variable properties
Bochanski et al. (2010): Luminosity and Mass functions of low-mass stars from SDSS
Boyajian et al. (2012): Temperature and radii of KM Dwarfs related to photometry
Curtis et al. (2019): example of stellar aging from rotation (i.e. variability)
Fantin et al. (2020): WDs in LSST
García–Berro & Oswalt (2016): White Dwarf luminosity function (LF)
Gizis & Stars (2021) : Brown Dwarf Cadence note
González Hernández & Bonifacio (2009): temperature of FGK dwarfs
Henry & McCarthy (1993): Mass Luminosity relationship GKM dwarfs
Medan et al. (2021): photometric metallicities of KM dwarfs
Ramírez & Meléndez (2005): photometric metallicity FGK dwarfs
Schmidt et al. (2016): Photometric metallicity/temperature of K/M Dwarfs

B.5.4. *The local IMF as inferred from nearby star forming regions and clusters*

Contributors: Luis M. Sarro

B.5.4.1. *Abstract* —Analysing the local Initial Mass Function and studying potential variations amongst various star forming regions requires a careful selection of members that is as consistent and homogeneous as possible in order to avoid biased determinations of this probability distribution. This has been done in the past based on hard cuts in proper motions and photometry, and more recently using probabilistic models of the distribution of sources (both the targets and the fore- and background contaminants) in astro-photometric space (including parallaxes when available). Including the effects of extinction and reddening has proved a difficult, computer intensive task.

So far, the incomplete and patchy data collected as part of the Dynamical Analysis of Nearby Clusters project (DANCe; Bouy et al. 2013) DANCe project contain $\sim 10^7$ sources with varying depth, amongst which at most only a few thousands are members. These 10^7 sources represent a lower limit to those expected from Rubin LSST. The LSST data set will include parallaxes (if available with sufficient accuracy), proper motions, and multi-band photometry, which can be used for membership determination using the Miec hierarchical bayesian inference code (Olivares et al. 2021).

Gaia astrometry reaches down to $G \sim 21$ which is insufficient to trace the low mass end of the IMF. Existing surveys suffer from severe inhomogeneities (both in depth, accuracy and completeness) which severely hinder the possibility to infer distributions, because the selection functions are often very poorly understood.

Examples and references for this kind of analysis can be found in Miret-Roig et al. (2022), Galli et al. (2021) and Olivares et al. (2021).

B.5.4.2. *Science Objectives* —

- Obtain complete censuses of nearby star forming regions of different ages in the LSST footprint
- Convert the existing photometry (Rubin LSST plus complementary external archives mainly in the near and mid infrared) into physical quantities (mainly effective temperatures and masses)
- infer probabilistically the Initial (or present-day) Mass Functions as a function of age for the SFRs in the Rubin LSST footprint.
- derive intermediate astrometric catalogs with improved uncertainties by using deep historic archival images, as exemplified by the Tycho-Gaia Astrometric Solution (TGAS) catalog.

B.5.4.3. *Challenges* —

- Prior to the Rubin LSST data releases, the main limitations are i) the unavailability of astrometry for faint sources and ii) the incompletenesses of the photometry. Existing data sets comprise a collection of data from very different sources and the elicitation

of the selection functions has been nearly impossible. As a result, the reliability of the inferred IMF is restricted to the mass ranges were completeness is achieved.

- The inference problem takes $\sim$ 10 hours for a selection of 10^5 likely candidate member sources in a dedicated computing facility with eight Nvidia GForce RTX 2080i GPUs. The preselection of these likely candidate members, however, is an undesirable step that leaves out sources with high extinction as well as outlying members, thus rendering the inferred IMFs incomplete.
- In principle, the quality of the astrometry will be the major impact factor affecting the results. The analysis would be carried out early in the release schedule and updated with subsequent data releases. New data will imply a better selection of members and a more reliable determination of the IMF for the low mass end which, depending on the star forming region and its extinction level, may be significantly undersampled in the first releases. For the bright end of the distribution, historic and archival data can be used to produce early good precision proper motions and parallaxes in a fashion similar to how the TGAS catalog was produced (Michalik et al. 2015).
- Extinction maps (either produced as part of the Rubin LSST data releases or external) will be needed as part of the membership analysis. In star forming regions with strong extinction, members can appear significantly displaced from the (isochronal) photometric sequence. Existing codes (Miec; Olivares et al. 2021) can (probabilistically) identify extincted members using preexisting extinction maps.

B.5.4.4. *Running on LSST Datasets (for the first 2 years)*—The first milestone for this project arrives with the availability of proper motions. As shown in the literature, proper motions are sufficient for complete (but contaminated) censuses. Parallaxes then allow (when and if available) for the cleaning of field contaminants with proper motions and photometry consistent with that of the SFR.

We propose to complement LSST positions with archival data from existing surveys and projects to produce long time baseline estimates of the proper motions that help constrain parallaxes that would be otherwise difficult to estimate with only 2 years data. According to Ivezić et al. (2019), proper motion uncertainties of the order of 0.2-1 mas yr$^-$1 (for r=21 and 24 respectively) are expected after 10 years of operation. After two years, the uncertainties would be a factor of 11 larger assuming a $t^{-3/2}$ scaling, with t being the time span of the observations. This implies proper motion uncertainties of the order of 11 mas yr^{-1} at r=24. For comparison, Miret-Roig et al. (2022) using archival data with a time baseline of 20 years in Upper Scorpio achieve average proper motion uncertainties of $\sim$ 1 mas yr^{-1} for $r < 24$. Having accurate proper motions derived using archival images can then allow for the determination of parallaxes (especially for nearby star forming regions where the signal is expected to be significant) in early data releases. For reference, parallax uncertainties of the order of 2.9 mas are expected after 10 years of operation which translates into 5.8 mas after two years (assuming a $t^{-1/2}$ scaling of the uncertainties). These uncertainties can be significantly reduced with the availability of accurate and precise proper motions.

B.5.4.5. *Precursor data sets*—Currently, the Gaia EDR3 provides the astronomy needed for the analysis only down to $G \sim 21$, which, depending on the youth and distance of the star forming region, translates into a mass lower limit that depends on the SFR. The Dynamical Analysis of Nearby Clusters project (DANCe; Bouy et al. 2013) provides the main stepping stone for testing the procedures. It provides photometry and proper motions based on historical images of a number of star forming regions and clusters. The data sets include typically tens of millions of sources but the inhomogeneity derived from combining $\sim 10^4 - 10^5$ images of very heterogeneous instruments, cameras and surveys translates into every complex selection functions and difficult to assess incompleteness.

B.5.4.6. *Analysis Workflow*—

1. For early data releases, derivation of astrometry using archival archival images for a long time baseline. This is a demanding computation for tens of thousands of images available before the start of operations of the LSST (see for example the analysis in Miret-Roig et al. 2022, for Upper Scorpio).
2. Selection in celestial coordinates of subsets of the catalog (one for each SFR). This includes proper motions, photometry and parallaxes when and if available.
3. Addition of complementary photometry if and when available (mainly near and mid infrared). This will require a cross match engine (either through the Rubin Science Platform or through an external service).
4. Addition of an extinction map, preferably in probabilistic form.
5. Identification of SFR members using Miec or evolutions thereof.
6. Probabilistic inference of IMFs by conversion of the isochronal photometric sequence into masses. Incorporation of Selection function.

B.5.4.7. *Software Capabilities Needed*—So far, Miec (Olivares et al. 2021) has only been applied to pre-filtered subsamples due to limitations in the computing power of existing infrastructure and the high computing needs of the software. This results in potentially incomplete censuses. As mentioned above, the existing software takes 10 hours to model a data set of 10^5 preselected sources in the Pleiades cluster (with negligible extinction) on an infrastucture with eight Nvidia GForce RTX 2080i GPUs. For the regions with high extinction that constitute the objective of this science case, we aim at computation times of several days using a much more powerful infrastucture and data sets of the order of millions.

There is no need for on-the-fly queries to the data base. All data (astrometry and photometry) can be queried prior to the application of Miec.

For the derivation of astrometric quantities using historic archival images, additional computing and storage facilities will need to be allocated. Typically, $10^4 - 10^5$ wide field images will be added to the LSST images.

If approved, the new astrometry can be made available to the community which would imply additional tables in the catalog.

B.5.4.8. *References for Further Reading* —Bouy et al. (2013): description of the Dynamical Analysis of Nearby Clusters project that illustrate the possibility to combine LSST images with historic archival images to derive improved astrometric solutions as in Michalik et al. (2015).

Olivares et al. (2021): description of the Miec hierarchical bayesian software to identify Star Forming Regions members and model their distribution in the space of astrometric and photometric measurements.

Miret-Roig et al. (2022): example of the kind of analysis proposed here as applied to the Upper Scorpio and ρ Ophiucus Star Forming Region.

B.6. *Local universe variable & transient science*

B.6.1. *Identifying symbiotic binaries by their color and variability*

Contributors: Gerardo Juan Manuel Luna (gjmluna@iafe.uba.ar)

B.6.1.1. *Abstract* —The deep, repeated and multi-filter observations during 10 years of the LSST provide an opportunity to search for the yet-predicted, but not found, population of 10^5 symbiotic binaries in our galaxy. Symbiotics are accreting binaries where a white dwarf (or a neutron star) accretes from the wind of a red giant. So far, about 400 symbiotics are known in our Galaxy and the Local Group, but theory predicts about 10^5 (Allen 1984; Magrini et al. 2003; Akras et al. 2019a). The luminosity in symbiotics can be powered either by accretion or by nuclear burning on the white dwarf surface. Shell-burning, luminous symbiotics, can be detected by searching for their strong Hα emission lines (Corradi et al. 2008, IPHAS). Those accretion powered, less luminous symbiotics, display flickering of significant amplitude (Luna et al. 2013). Color-color and color-variability diagrams from surveys such as SkyMapper (Figure 3), WISE and Two-Micron All Sky Survey (2MASS) have been used to identify regions where symbiotics might stand-out. This proposal aims to use LSST color-color and color-variability amplitude diagrams in order to identify new symbiotics either in our Galaxy or within the Local Group.

B.6.1.2. *Science Objectives* —

- Identify new symbiotic star candidates using color-color diagrams and variability information.
- Further spectroscopy follow-up to nail-down the true symbiotic nature of the candidates

B.6.1.3. *Challenges (what makes it hard)* —

- Identify which combination of LSST filters is most suited to discriminate symbiotics from other objects such as cataclysmic variables, T Tauri, planetary nebulae, Young Stellar Object (YSO).
- Identify variability in the light curves of different filters.
- Build the control sample of known symbiotic stars observed with the LSST filters in order to determine where most symbiotics are located in the color-color diagram. An alternative to obtain this control sample before the start of the LSST is to use archival optical spectra and obtain colors in the LSST filters system.
- Implement machine-learning techniques to automatize the identification process and determine the regions on the color-color diagram where most symbiotics would be located (Akras et al. 2019b).

B.6.1.4. *Running on LSST Datasets (for the first 2 years)* —

- We will analyze both the alert stream and the data release light curves, incorporating them to the best color-color and color-variability diagrams.

- Once a candidate is identified, we will take an optical spectrum in order to verify its symbiotic nature.

B.6.1.5. *Precursor data sets* —

- Optical spectra from approximately 100 symbiotic in our own archive plus spectra taken by amateurs and available in the Astronomical Ring for Access to Spectroscopy (ARAS) database

B.6.1.6. *Analysis Workflow* —

- Remove poorly calibrated photometric data and sources flagged with suspicious photometry (e.g. on edge of Charge-Coupled Device (CCD) or diffractions spike).
- Remove/flag outlier measurements from a light curve
- Remove/flag extended sources
- Populate the color-color and color-variability diagrams.
- Data storage and archives
  - Collect control samples and store in a DataBase (DB)
- Variability Identification
  - Characterize variability amplitude (e.g. rms of light curve) and filter based on this amplitude

B.6.1.7. *Software Capabilities Needed* —

- Ability to apply selection filters to data (SQL query)
- Storage of light curves as objects (time, passband, noise, flags for bad points) with annotations (e.g. classifications) that can be queried and cross matched to other data sets
- Storage of outputs of filtered and classified source candidates
- Visualization of postage stamps and light curves for individual sources

B.6.1.8. *References for Further Reading* —

- Akras et al. (2019b)
- Lucy (2021)

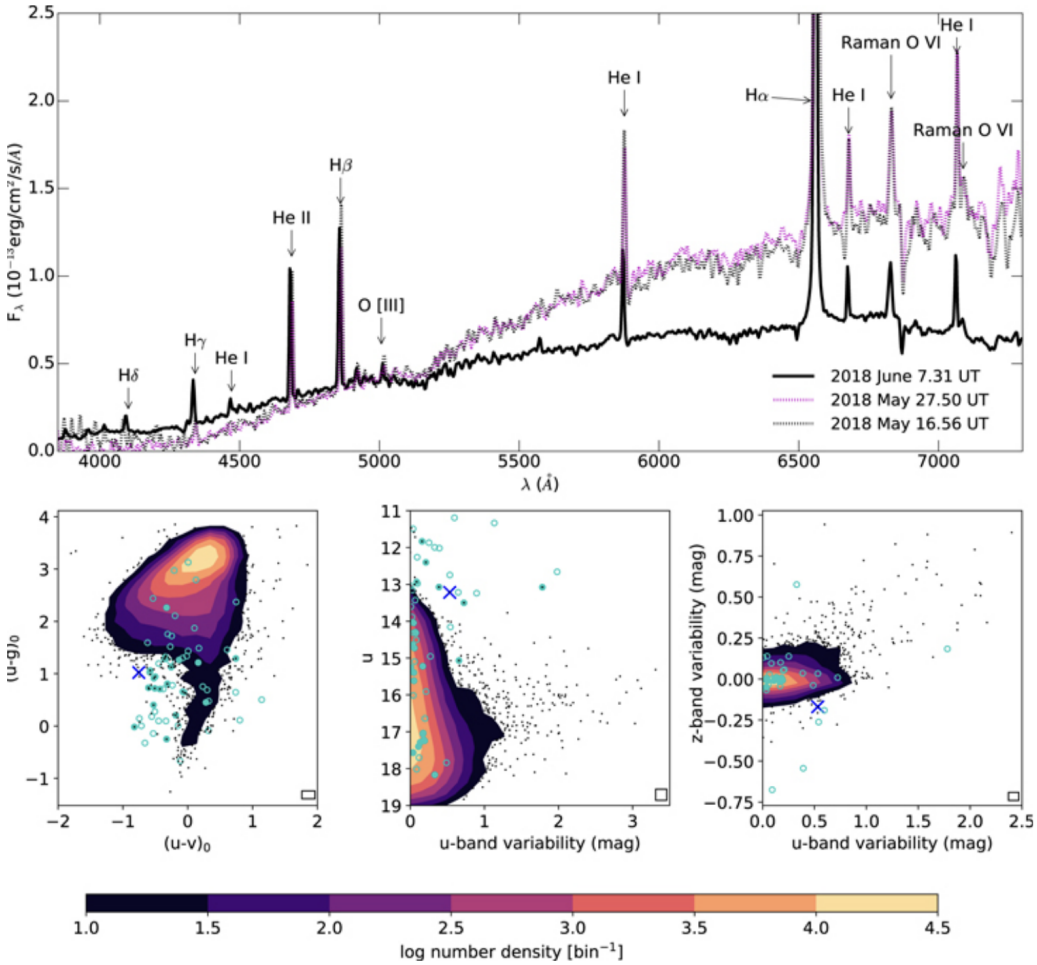

Figure 3. Upper panel: Our follow-up spectra of Hen 3-1768 exhibited the labeled emission lines. The first two were obtained with a 279mm Schmidt-Cassegrain telescope equipped with a LISA spectrograph (R~1500-1900, smoothed here to R-700; dotted lines) on 2018 May 16.56 and 27.50 UT and reduced in ISIS, the third at CASLEO with the REOSC spectrograph (R-700; solid line) on 2018 June 7.31 UT and reduced in IRAF. Flux calibrations are extremely preliminary, and absolute flux calibration was unavailable for the 2018 May 27.50 spectrum. Lower panels: We selected Hen 3-1768 (blue cross) from a sample of ~210,000 RGs (black points with contours over dense regions), because known symbiotics (turquoise circles, filled for symbiotics in the RG sample and hollow for symbiotics excluded from the RG sample by IR or quality cuts) are outliers in spaces defined by these parameters from SkyMapper data: (u-g)0, (u-v)0, average u magnitude, maximum u variability between any two SkyMapper epochs, and z variability between those same two epochs (available for only ~90,000 RGs). The densest regions of parameter space are replaced by contours delineating the logarithmic number density of RGs (per bin; hollow rectangles). De-reddening was performed using total Galactic extinctions. Intervals between epochs range from 3 minutes to 1.5 years; the plotted interval for Hen 3-1768 is 142 days, larger than the sample median of 15 days. (Figure credit: Lucy et al. 2018)

B.6.2. *Light Echoes: study the reflection of transients on interstellar medium in the LSST Era*

Contributors: Xiaolong Li (lixl@udel.edu), Federica Bianco (fbianco@udel.edu)

B.6.2.1. *Abstract*—Light Echoes (LEs) are the reflections of astrophysical transients on interstellar dust. LEs enable the study of the dust as well as the source transients. But they

are rare and extremely difficult to detect because they appear as faint, diffusive features, presenting along with a lot of false positives, such as the star streaks, reflections of telescope, satellites, artifacts and so on. To date, only a few examples (~100) have been discovered.

While LSST will detect thousands of transients in the sky and release millions of alerts every night, the Rubin data processing pipeline is designed for point-sources, and it will entirely miss diffuse transient features (in the low SNR regime).

The combination of Rubin LSST's cadence, depth, and excellent image quality will have the potential to revolutionize LEs studies, pushing it from serendipity discovery to a statistical regime. But this requires the creation of data processing and analysis software tools specific for this purpose starting at the image-analysis level. While Rubin's WFD strategy with repeat visits every ~ 3 days would be excellent for LE studies, LEs appear in dusty regions where the LSST observing strategy may be different than the nominal WFD, potentially requiring imaging follow-up observations to be conducted at other observatories. Spectroscopic follow-up is critical to characterize the source transient (e.g., spectral typing, temperature constraints). To maximize the science throughout, an end-to-end pipeline is necessary. This pipeline should include modules for the detection of the LEs, built upon LSST's image data, for automated follow-up management (including selection of the best candidates to follow up and integration with telescope observing software), and follow-up spectroscopic analysis in order to characterize the transients and constraint the dust properties.

B.6.2.2. *Science Objectives* —

- Study the evolution of LEs over time in the LSST image series
- Analysis of the LE spectra to characterize the transient
- Reconstruct the relevant dust properties (geometry, density)

B.6.2.3. *Challenges* —The end-to-end LE pipeline would include:

- Software to detect LEs from LSST's image data within ~days-to-weeks from LE first brightening to enable spectroscopic follow up before the LE vanishes;
- Software to automate follow-up spectroscopic (and possibly imaging) observations
- Software for the identification of the source transients through multi-epoch directional analysis and cross-matching with historical catalogs;
- Pipelines for the analysis of imaging and spectroscopic data to constrain properties of the reflecting dust and emitting transient.

Challenges include

- Potential need to tune the Rubin image subtraction models to work effectively in the low SNR regime for diffuse sources,
- Creating a robust ML/Artificial Intelligence (AI)-based detection model that can handle the large volume of LSST image data and has high recall (as the phenomenon is rare) and high precision (as there are many potentital sources of false positives) is complicated by the small number of LEs to train the model on,

- Science throughput is enhanced by cross-matching and joint analysis of multi-band observations including infrared, X-ray, HI, and CO to study the reflecting interstellar medium properties, with the usual challenges associated to cross-matching that are discussed elsewhere in this document.

B.6.2.4. *Running on LSST data sets (for the first 2 years)*—For the purpose of searching for LEs, accurate sky templates in dusty regions of the sky are critical. While the Rubin-DM-pipeline-generated templates may be sufficient for our purposes, the current plan to minimize observations in dusty regions suggests the pipeline may not be specialized to handle these data and the template generation process may need revisions. We may need to re-build template, especially dust regions. As the Rubin detection pipeline will not discover extended transients at the SNR limit we will need to run a specific detection pipeline for LEs. If we are unable to reprocess all LSST images we would prioritize regions rich in interstellar dust, which have a higher probabilities of reflecting light of transients.

We'll create LE catalogs that include images postage stamps and transient sources. The postage stamp size will vary depending on the distance, size, and angle of the reflecting dust as LEs can get as large as $\sim 1'$

B.6.2.5. *Precursor data sets*—Precursor datasets include ATLAS and DES (and other DECam data). Both datasets are currently being used to create training sets for LEs discovery. Images from DECam, while shallower, have comparable quality with LSST; a cross analysis between LSST and DECam needs to be developed.

B.6.2.6. *Analysis Workflow*—

- Image preprocessing
- Creating templates;
- Image subtraction: derive the subtracted images
- Detection model: input images to detection pipeline
- Prioritize detections for follow up
- Identify sources through photometric and spectroscopic analysis
- Cross analysis with other passband observations
- Retrieve relevant interstellar dust structures

B.6.2.7. *Software Capabilities Needed*—

- Query LSST image data, including position and image specialization;
- Access to the interstellar dust map;
- Able to deploy well-trained detection models;
- A visualization portal that can easily view Flexible Image Transport System (FITS) images, annotating and classifying;
- Analysis tools for photometric and spectroscopic data

B.6.2.8. *References*—Rest et al. (2012): Light Echoes of Transients and Variables in the Local Universe.

Patat (2005): Reflections on reflexions I. Light echoes in Type Ia supernovae.
Patat et al. (2006): Reflections on reflexions II. Effects of light echoes on the luminosity and spectra of Type Ia supernovae.

B.6.3. *Compact White Dwarf Binaries in LSST*

Contributors: Alekzander Kosakowski (alekzander.kosakowski@ttu.edua)

B.6.3.1. *Abstract*—Double-degenerate compact white dwarf binaries will be the dominant source of gravitational wave emission detectable by the upcoming Laser Interferometer Space Antenna (LISA) mission. On the order of $O(10^2)$ systems are predicted to show measurable variations in both gravitational waves and electromagnetic radiation (Korol et al. 2017). Large sky surveys such as LSST allow for the efficient identification of photometric variables in anticipation of the launch of LISA. Compact white dwarf binaries may show variability due to eclipses, tidal distortions, relativistic beaming, and reflection. These sources of variability are regularly used to characterize the physical properties of the binary, providing clues to its formation and eventual fate.

The Zwicky Transient Facility (Bellm et al. 2019; Masci et al. 2019) has enabled the swift discovery of such binaries in the northern sky. Simple period-finding searches such as Lomb-Scargle (Lomb 1976; Scargle 1982), conditional entropy (Graham et al. 2013), and Box Least Squares (Kovács et al. 2002) have resulted in many scientifically valuable discoveries (see Burdge et al. 2020). We expand this search to the southern sky using LSST and BlackGEM to create a catalog of photometrically variable, compact, double-degenerate binaries that will merge due to the emission of gravitational waves. Combined with the ELM Survey (Brown et al. 2020; Kosakowski et al. 2020), which detects binarity through a photometric and astrometric selection followed by measured radial velocity variability, we aim to create a complete catalog of compact white dwarf binaries that can be used to provide clues to the relatively-poorly understood stages of compact binary evolution, such as common-envelope evolution, and the formation rates of Type Ia supernovae and post-merger helium-rich objects. Having a complete list of compact binaries will provide valuable insight to the expectations of LISA. The combined data sets from various surveys over many years additionally facilitates direct measurements of the effects of gravitational waves (and tidal effects) on the orbits of these compact binaries (see Hermes et al. 2012).

B.6.3.2. *Science Objectives*—The main objective of this project is to create a catalog of photometrically-variable white dwarfs in the southern sky, with strong emphasis on ultra-compact ($P \lesssim 60$ min) white dwarf binaries, which will act as LISA verification sources. Because this is a large sky survey, many other variable white dwarfs will be discovered, classified, and cataloged, including massive rotating white dwarfs, pulsating white dwarfs, and various eclipsing white dwarf binaries.

Follow-up spectroscopy will be performed on candidate binaries to confirm their spectral classification, atmospheric parameters ($T_{\rm eff}$ and $\log g$), and the presence of a companion evidenced by radial velocity variations.

- Crossmatching with other surveys such as Gaia, VST ATLAS, SkyMapper, and BlackGEM will be the first step to identify the relatively bright white dwarf binaries. Fainter white dwarfs only visible to LSST will be identified through photometry and astrometry cuts to the LSST data as it becomes available.

  - Even a simple positional cross-match can be slow when using large surveys such as Gaia. Developing a parallelized cross-matching algorithm will be necessary to efficiently identify targets across many surveys.

- Light curve processing may need to be completed on the user's local machine, thus a form of bulk data download is necessary in order to allow users to work off of the Rubin Science Platform for more advanced analysis such as light curve modeling.
  - Remotely processing many light curves may be too taxing on the Rubin Science Platform servers considering that the resources are shared.
  - This will require a streamlined command-line tool for users to easily query for large numbers of sources at once.
- Use of the RSP to run period finding algorithms on a large number of light curves at once is important. Many useful period finding algorithms exist and can be ready for bulk use after minor modifications.
- Finally, classifying each light curve based on features identified in their light curves (periodicity, amplitude, variability shape, eclipse duration, etc) will allow for an organized catalog of white dwarf variables.
  - This objective will require a machine learning classification algorithm. Use of existing data from other surveys will help create training sets for classification.

B.6.3.3. *Challenges (what makes it hard)* —

- Because white dwarfs are inherently faint due to their size, nearby field stars may easily dominate a blended Point Spread Function (PSF) in crowded fields. Developing a deblending algorithm capable of handling a large difference in stellar PSF profiles while not significantly affecting potentially weak periodic signals is a non-trivial task.
- Period-finding algorithms on dense frequency grids are computationally expensive, especially when trying to properly handle multi-band data. Developing an efficient algorithm capable of identifying multiple types of variability and handling multiple filters simultaneously is not an easy task.
- The algorithm will need to be rerun as more data is taken. Because computational resources are likely to be scarce, determining the most efficient times to rerun the periodogram will also be a challenge.
- Follow-up time-series spectroscopy is expensive.

B.6.3.4. *Running on LSST Datasets (for the first 2 years)* —

- This project will make use of the calibrated light curves and FITS images taken over the first two years in combination with other large survey data.
- LSST light curves will be combined with other surveys, such as ZTF and BlackGEM to increase temporal sampling and allow for rapid identification.
- Focused deep drilling fields may provide early detection of scientifically valuable binaries.

- Given the expected eclipse durations for these double-degenerate binaries (~60 s), having greater than ~100 data points across all filters may allow eclipsing binaries to be identified programatically. Sinusoidally varying systems may be identified with fewer data points.

B.6.3.5. *Precursor data sets* —

- ZTF's southern-sky overlap with LSST will provide a valuable test-case for our analysis.
- During the early stages of LSST, SkyMapper, VST ATLAS, Gaia DR3, and BlackGEM will provide valuable information that will augment the LSST data.

B.6.3.6. *Analysis Workflow* —

- "Poor quality" images will removed using the recommended LSST data quality flags.
- For the relatively bright objects still visible in other surveys, the "good quality" LSST light curve data will be combined with light curve data from other surveys to produce more complete light curves with longer baselines and better sampling than LSST alone would provide.
- Various periodograms will be run on each light curve to identify different types of variability. Initially, a simple Lomb-Scargle and Box Least Squares periodogram can be run to identify compact ellipsoidal and eclipsing binaries.
  - Light curves that do not show M N-sigma deviant points through LSST alerts should not use resources on an expensive Box Least Squares search for eclipses. Determining M and N such that we efficiently identify real eclipses while successfully rejecting false positives is a challenge; too few cuts will result in significantly higher run time and time lost filtering false positives from the final sample, while too many cuts will result in many missed eclipsing binaries. It is important to use forced-photometry for these light curves and not reject truly deep eclipses from the light curves.
- Create phase-folded light curves at the most-probable frequency determined through each periodogram.
  - For Lomb Scargle: record light curve statistics of the phased light curve, including most-probable period, amplitude of variability, and goodness of fit for a single- and double-sine fit.
  - For Box Least Squares: record most-probable period, eclipse depth, eclipse duration, and number of in-eclipse data points per filter.

B.6.3.7. *Software Capabilities Needed* —Given the volume of data expected to come from the LSST program, performing a complete analysis on even subsamples of this data will be computationally expensive. White dwarfs in general display a broad range of photometric variability in terms of shape and period, with variations between milli-magnitudes and

magnitudes on periods on the order of single-minutes to days. To perform large-scale periodicity searches on candidate white dwarfs requires an efficient algorithm capable of equally identifying both short- and long-period eclipsing systems (such as Box Least Squares) and sinusoidally-varying systems (such as Lomb-Scargle and Conditional Entropy). However, the scale of this problem may be reduced if only considering compact gravitational wave LISA verification binaries, which have a relatively narrow frequency range.

- The efficient completion of this project requires a consistent object identifier number across all filters. Having a common multi-filter identifier allows for quick extraction of all epochs per object and allows each object to be assigned complete LSST color information which will be used for astrometric and photometric target selection of candidate binaries. Astronomers working on specific objects looking to crossmatch their work with LSST will greatly benefit from a bulk light curve search on position that returns a single object ID per object.
- A sophisticated period-finding algorithm will need to be run in order to identify variability and perform classification. Even restricting our period search to within the expected detection limit on LISA, the volume of data expected requires a highly-parallelized algorithm capable of handling non-regular light curve sampling and heteroskedastic flux errors, which would be run on a state-of-the-art high-performance computing center. A GPU-accelerated program may perform the periodicity search over a dense grid best. However, over the 10-year LSST baseline, many of these LISA verification binaries will show measurable period decay, which may negatively affect a standard periodogram's ability to detect orbital periods. As the LSST project evolves, the need for a periodogram that detects both period and period decay will increase.
- The results may be presented to the user through an online GUI. Given the users target selection region, the back-end would collect target information, such as colors, light curves, and periodograms, and present it in a table format. Selecting specific targets on the table would display the calibrated LSST light curve, its periodogram, the calibrated light curve phase-folded to the most-probable frequency and its half-frequency, and the object's location on the HR diagram for quick by-eye classification. The calibrated LSST light curve will help eliminate flaring systems, such as Cataclysmic Variables. Providing a phase-folded light curve at the half-frequency will help classify ellipsoidal variability and reflection effects.
- Combining light curves from LSST with other southern-sky surveys may provide significantly better sampling, which would improve the likelihood that periodic variables are detected at their true periods and increase detection efficiency. With this in mind, the LSST program would benefit from providing the user the ability to import their own light curve data and augment the LSST data for additional periodogram analysis.

B.6.3.8. *References for Further Reading*—Bellm et al. (2019), Brown et al. (2020), Burdge et al. (2020), Kovács et al. (2002), Graham et al. (2013), Hermes et al. (2012), Korol et al. (2017), Kosakowski et al. (2020), Lomb (1976), Masci et al. (2019), Scargle (1982)

B.6.4. *Analysis of Microlensing events by stars and compact objects*

Contributors: Rachel Street (rstreet@lco.global), and TVS Microlensing Group

B.6.4.1. *Abstract*—The technique of microlensing is routinely used in the Galactic Plane to explore populations that are too faint for direct electromagnetic detection. Requiring only multiband optical time series photometry of a background source star for detection, rather than of the target lens that passes in front of it, microlensing will provide insights into the population of low mass stars, planets and isolated compact objects in regions across the Milky Way where they are otherwise inaccessible. Historically, almost all lensing events have been identified in the Galactic Bulge; LSST will discover thousands of events per year across the Galactic Plane, allowing us to explore these populations in a greater range of formation and evolutionary contexts. Lensing by isolated black holes will enable us to constrain the black hole mass function and hence theories of their formation (e.g., Gould 2000). This and other complementary science is described in Street et al. (2018a) and Poleski & Mróz (2018). Furthermore, LSST will operate contemporaneously with the Roman Space Telescope's near-infrared survey of the Galactic Bulge, and can provide highly complementary optical data that will not only provide additional model constraints for >1400 bound exoplanetary events, but also constrain the masses of ~250 free-floating planets, and increase the number of planetary anomalies discovered by filling in gaps in the Roman survey cadence (Street et al. 2018b).

Models of microlensing are derived from the multi-band lightcurves, but events suffer from multiple degeneracies, so the analysis of anomalous events in particular entails searching a large and non-linear parameter space for the best fit. With LSST expected to identify thousands of events, but provide relatively low cadence data in most cases, this model fitting process will be computationally intensive, but highly parallelizable, as each event can be evaluated independently. For some events where contemporaneous lightcurve data is available from other facilities, the analysis will need to access external data catalogs (e.g., from Roman) and include these data in the model fitting process.

B.6.4.2. *Science Objectives*—

- Evaluate how well microlensing events are identified from the LSST alert stream by different brokers and the insights gained by combining their output. Tests (e.g., Godines et al. 2019; Mróz et al. 2020) have shown that classification algorithms are most effective when a baseline lightcurve of ~1 yr duration is available, with a photometric precision of <1%.
- Identify microlensing anomalies from the lightcurve data, ideally while the event is ongoing and the lightcurve data is incomplete. Prior work exists in this area (see above links, e.g.) but needs implementation in an LSST context.
- Implement parallelized modeling of large numbers of lensing events. This will need to robustly identify cases where degeneracies or data gaps warrant further evaluation. Several open-source modeling packages are available (listed on the Microlensing-

Source website Microlensing Software Developers 2019, see Bachelet et al. 2017; Poleski & Yee 2019; Bozza 2010).

B.6.4.3. *Challenges (what makes it hard)* —

- The scale of the analysis: infrastructure is needed to manage the modeling of (potentially) hundreds of events at any given time, while keeping track of the status of each one.
- Identifying events that need additional attention: some events can be accurately modeled in a fully automated manner, but many suffer from model degeneracies, especially for the scientifically most interesting events.
- Identifying anomalies in real-time, to enable follow-up observations to be triggered where appropriate. Ideally this should take place with as short a latency as possible, within hours. Modeling of ongoing events will need to take place continuously, with updates whenever new data are acquired.
- Reliably combining downstream alert streams from different brokers.
- The cadence of LSST lightcurves, depending on the observing strategy. Combining data from multiple passbands and acquiring follow-up data where necessary can help to address this.

B.6.4.4. *Running on LSST Datasets (for the first 2 years)* —This science will utilize the real-time LSST alert stream and lightcurve catalogs. For effective false-positive rejection, we anticipate at least 6–12 months of baseline data will be needed. It is expected that (approximately) thousands of events will be detected per year (Sajadian & Poleski 2019). The derived data product will be a catalog of the fitted model parameters and derived physical parameters for the sample of lensing events.

B.6.4.5. *Precursor data sets* —There are various datasets from existing surveys that are currently being analyzed for microlensing events, some of which are public, for example, some ZTF data, Gaia data, MOA data.

B.6.4.6. *Analysis Workflow* —

- Query multiple brokers for microlensing detections and combine their assessments to prioritize events.
- Fit multi-band LSST lightcurves together with any other lightcurves available, in real-time.
- Review combined lightcurves for anomalous deviations; if detected, evaluate for follow-up potential.
- Identify and evaluate events subject to degeneracies.
- Analyze the CMD for the region of the event and constrain source parameters.
- Derive model, and hence physical, event parameters

B.6.4.7. *Software Capabilities Needed* —

- Detection and accurate classification of microlensing events in progress by LSST alert brokers, as well as access to the resulting alert data. A number of algorithms are in place at ANTARES and other brokers.
- Real-time access to LSST lightcurve data on (RA, Dec) query, and previous Data Release photometry for the target and surrounding region of sky.
- Access to Roman lightcurve catalogs on (RA, Dec) query, as well as Gaia, VVV and other data catalogs for brighter events. This should be available from a NASA data archive.
- Infrastructure to automate and parallelize the modeling of hundreds of events simultaneously and access to a Central Processing Unit (CPU) cluster facility on which to run the model processes. GUI to enable user monitoring of the modeling processes
- Real-time anomaly detection software. Prior algorithms are available but will need revamping for LSST's context.
- TOM system to manage follow-up observations. A prototype system is in operation, which could also manage the modeling processes if deployed suitably.
- Multi-TB storage for lightcurve collections and TOM/modeling system database.
- Webservice to serve GUI for TOM system to collaborators worldwide.
- Note that the current state of the art for the modeling of a single binary lensing event is ~2 hr.

B.6.4.8. *References*—Gould (2000): A Natural Formalism for Microlensing
Street et al. (2018a): The Diverse Science Return from a Wide-Area Survey of the Galactic Plane
Poleski & Mróz (2018): The First Extragalactic Exoplanets — What We Gain From High Cadence Observations of the Small Magellanic Cloud?
Street et al. (2018b): Unique Science from a Coordinated LSST-WFIRST Survey of the Galactic Bulge
Godines et al. (2019): A machine learning classifier for microlensing in wide-field surveys
Mróz et al. (2020): Gravitational Microlensing Events from the First Year of the Northern Galactic Plane Survey by the Zwicky Transient Facility
Microlensing Software Developers (2019): Microlensing Source - Publically available software for microlensing modeling, simulation and analysis
Bachelet et al. (2017): pyLIMA: An Open-source Package for Microlensing Modeling. I. Presentation of the Software and Analysis of Single-lens Models
Poleski & Yee (2019): Modeling microlensing events with MulensModel
Bozza (2010): Microlensing with an advanced contour integration algorithm: Green's theorem to third order, error control, optimal sampling and limb darkening
Sajadian & Poleski (2019): Predictions for the Detection and Characterization of Galactic Disk Microlensing Events by LSST

B.6.5. *Young stellar objects and their variability*

Contributors: Sara (Rosaria) Bonito, (rosaria.bonito@inaf.it), Rachel Street, Sabina Ustamujic (sabina.ustamujic@inaf.it), Laura Venuti (lvenuti@seti.org)

B.6.5.1. *Abstract*—YSOs show short-term as well as long-term photometric variability related to the physical processes at work in these complex systems and their geometry, e.g. mass accretion from circumstellar disks (which can proceed in a steady, funnel-flow pattern as well as in bursts or eruptive events), flares, rotation, or the presence of dusty warps within the inner disks. Monitoring the variability over different timescales is critical to constrain the dynamics of these processes, and studying the color dependence of such variability is essential to disentangle the characteristic signatures of distinct potential mechanisms at play. Vera C. Rubin Observatory LSST will be the ideal instrument to allow us to obtain well-sampled multicolor lightcurves of star forming regions (as e.g. Carina, Orion Nebula Cluster, NGC 2264, NGC 6530, NGC 6611) to acquire the first statistically significant data on how young stars vary on both short and long timescales. Thanks to its depth and duration, the Rubin LSST survey will provide us with crucial information to both discover new populations of accreting young stars for classification, and to characterize known objects. From the point source catalogs extracted from stacked images, we will be able to achieve the most extensive reconstruction to date of the manifold accretion processes at play in YSOs, and of how they evolve as a function of stellar mass and average cluster age. A deep, uniform sky coverage is crucial to ensure statistical representation for young stars of all spectral types in clusters, and to sample the impact of different cluster ages and different environmental conditions (e.g., field crowdedness, presence of massive stars). By exploring the light curves acquired with LSST at different filters (u,g,r,i), we can characterize and classify YSO variability. We can discriminate different processes at work in YSOs by examining how the location of individual sources changes with time on color-color and color-magnitude diagrams. From the comparison between individual stellar colors and reference photospheric color sequences, we can discern accreting young stars, or Classical T Tauri starss (CTTSs), from non-accreting young stars, or Weak-lined T Tauri starss (WTTSs), while the spectrum of the observed flux variations (i.e., amplitude of color variations and slope of photometric variations on color-magnitude diagrams) on different timescales can reveal whether such variations are dominated by accretion events or circumstellar extinction events (e.g., by inner disk warps). See White Paper by Bonito et al. (2018), Cadence Note by Bonito et al. (2021), and Venuti et al. (2014, 2015) for more information on the science case discussed here.

B.6.5.2. *Science Objectives*—

- Variability due to different physical processes in young stellar objects
- Discrimination of different processes at work in young stars
- Description of both short-term and long term variability in young stellar objects
- Classification and characterization of variable stars
- Spectroscopic follow-up of variable processes as, e.g., EXor objects (Herbig 2008)

- We need a proper cadence to populate the lightcurves to follow short-scale variability
- We require the bluest filters to describe the accretion process in YSOs
- Multi-filter observations will allow us to explore color-color and color-magnitude diagrams of YSOs

B.6.5.3. *Challenges (what makes it hard)* —

- A proper cadence to retrieve different shapes of lightcurves for different processes at work is needed for the short-term variability
- Classification of different kind of variability
- Alert stream for variability of EXor objects (long-term variability) is important also for spectroscopic follow-up
- Early Science exploration of a testbed star forming region (e.g. Carina) is strongly suggested to extend the investigation to other SFRs (see Bonito et al. 2018).

B.6.5.4. *Running on LSST Datasets (for the first 2 years)* —

- Early Science exploration of Light Curves (LCs) in one selected star forming region (e.g., Carina) has been proposed in Bonito et al. (2021) as a micro-survey for the optimization of the survey strategy
- Similar micro-surveys can extend the investigation to other SFRs (see Bonito et al. 2018): we have identified as ideal a coverage of the LC with 140 points in 1 week for each filter:
- one point every 30 minutes in a 10 hour/night observation for 7 consecutive nights,
- motivated by the fact that this sampling will allow us to reveal short-lived phenomena in
- YSOs with a probability of 5% (Fig. 1, Bonito et al. 2021)
- Additional spectroscopic follow-up and multi-band (including also UV and X-ray data) will complete the characterization of the sample

B.6.5.5. *Precursor data sets* —

- We have simulated the Rubin LSST LCs cadence by using Convection, Rotation et Transits planétaires (CoRoT) and K2 data for short-term variability and ZTF data for EXor objects (long-term variability)
- Earlier large-scale surveys covering the same area of the sky to be targeted with LSST in similar filters (e.g., ZTF, VPHAS+, Drew et al. 2014; SkyMapper Southern Sky Survey, Wolf et al. 2018; PanSTARRS, Chambers et al. 2016) will be ingested and will serve both as preliminary tools for the identification of a reference sample of well-known young stellar populations in the region, and to extend the time coverage of the LCs reconstructed for our LSST targets.
- Photometric cross-calibration between different surveys will be achieved statistically by selecting the population of field stars (not YSOs, which are intrinsically variable) common to all catalogs and spatially located in the same region as our young stellar populations of interest (see, e.g., Venuti et al. 2014).

B.6.5.6. *Analysis Workflow* —

- Preliminary mining of data archives from large-scale surveys available in the literature at similar wavelengths (e.g., VPHAS+, SkyMapper Southern Sky Survey), to inform an initial catalog of known young stellar populations in the areas to be surveyed with LSST.
- Analysis of diagnostic diagrams, such as the distribution of photometric scatter vs. apparent magnitude (e.g., Venuti et al. 2021) and the distribution of flux aperture radius vs. apparent magnitude on the CCD frame, to flag and discard saturated data, spurious detections (e.g., cosmic rays), and background-dominated point sources.
- Identification of young stellar objects in the LSST catalog, by inspecting properties such as photometric clustering on color-magnitude diagrams (to locate the sequence traced by cluster stars over the field population), and kinematic or astrometric association in the field. The preliminary catalogs of young stellar populations built from the literature mining step will be used as a guide to interpret the distributions in photometric properties of LSST targets. This step will enable discoveries of new young stellar populations, as well as the achievement of a more complete census for already known young stellar populations, thanks to the unprecedented depth of the LSST survey (critical, for instance, for studies of the stellar initial mass function).
- Cross-correlation of different catalogs based upon the RA, Dec coordinates (using tools such as Tool for OPerations on Catalogues And Tables (TOPCAT); Taylor 2005) to match the same sources detected with different filters, at different epochs, or from different surveys, and thereby reconstruct the light curve of each identified young star. The matching radius will be defined based on the astrometric precision of each catalog, and a spatial analysis of field crowdedness (or typical intra-source separation in the region under exam) will be conducted to assess the probability of fortuitous matches resulting from the cross-correlation procedure.
- Analysis of color-color and color-magnitude diagrams to separate young non-accreting stars (which exhibit colors consistent with photospheric/chromospheric emission) from young accreting stars (which exhibit continuum excess emission compared to the photospheric level as a result of the emission from the accretion shocks; revealed particularly at short wavelengths, yielding bluer colors compared to non-accreting young stars in the same population; e.g., Venuti et al. 2014). The same analysis will be repeated for all single-epoch catalogs, and the frequency of cases that switch between accreting/non-accreting classifications at different epochs will provide critical information on potential transient accretion states during the later stages of protoplanetary disk evolution (e.g., Cieza et al. 2013).
- Implementation of multiwavelength variability indicators (e.g., Stetson 1996) to identify young stellar objects that exhibit significant variability above the noise level traced by field stars.
- Implementation of previously tested techniques, such as studying the color slopes associated with distinct young stellar variables (e.g., Venuti et al. 2015; Hillenbrand

et al. 2022), and assessing the degree of periodicity and asymmetry in the measured flux variations (e.g., Cody et al. 2014; Cody & Hillenbrand 2018; Hillenbrand et al. 2022), to discern the dominant physical drivers of the observed variability behaviors (e.g., geometric modulation by surface spots at a given temperature; stable magnetospheric accretion activity vs. episodic accretion; circumstellar occultation).

- Analysis of the wavelength-dependent amplitudes of variability measured as a function of epoch difference (from hours, to days, to years), to parse the dominant timescales of variability for each photometric behavior (e.g., Sergison et al. 2020), and the intensity of flux variations (as a fraction of the typical luminosity state of the object) that are triggered over those timescales.
- Identification of the most reliable predictors of young stellar status and variability behavior (i.e., specific color signatures, specific time intervals over which distinct variability processes can be most easily distinguished) and development of automated routines to extend the same classification analysis from the testbed case (microsurvey of the Carina star-forming region) to other regions encompassed by the LSST survey. This step will be crucial to assess the time evolution of the physical processes that govern the star-disk interaction over the protoplanetary disk lifetimes.

B.6.5.7. *Software Capabilities Needed* —

- *ugri* bands for LC, color-color, and color-magnitude diagram analysis
- Alert Brokers for long-term variability and spectroscopic follow-up is important for EXor objects
- We plan to develop a software for the classification of YSO variability at all the time-scales
- Interactive visualization will be used
- We plan to use Three-dimensional (3D) models and 3D rendering to reproduce the configurations that lead to the observed LCs. The use of 3D models will allow to explore different line of sights and to understand how geometric effects affect the observed LCs

B.6.6. *Long Period M dwarf Variability*

Contributors: Mark Popinchalk (popinchalkmark@gmail.com)

B.6.6.1. *Abstract* —M Dwarfs are the most common type of star, but are difficult to observe due to their intrinsic low luminosities. They are also the most long-lived stars, making them important for understanding the oldest limits of gyrochronology. Previous M dwarf studies include the MEarth sample (Newton et al. 2018), which looked at a volume limited sample of 25 pc and successfully recovered rotations for objects down to Gaia G $\approx$ 17 mag. These M dwarfs have a roughly bimodal distribution of periods, with peaks at $\sim$1 and $\sim$100 d and relative sparsity in between. This points to a potential sudden transition from rapid to long periods, but the age at which this occurs is poorly constrained. The cause of the spin down is also unknown. Furthermore, not all M dwarfs are the same, as there is a transition from partially convective to fully convective interiors around the M3 spectral type. This is thought to have implications for angular momentum evolution, as descriptions used for core and envelope angular momentum transfer in more massive stars is not applicable to fully convective objects. Current space based missions like Transiting Exoplanet Survey Satellite (TESS) struggle to 1) measure periods >28 days, and 2) recover periods from objects fainter than Gaia G = 17 mag. The depth and duration of LSST will be a powerful tool for increasing the number of M dwarfs with rotation periods $\gtrsim$30 d.

B.6.6.2. *Science Objectives* —

- Identify M dwarfs in LSST.
- Measure the rotation periods for a large sample of M dwarfs.
- Probe the distribution of long-period rotation in M dwarfs, and understand the importance of the transition from partially to fully convective interiors.
- Constraining the evolution of angular momentum in M dwarfs.
- Investigating color dependence in M dwarf variability.

B.6.6.3. *Challenges (what makes it hard)* —

- Because the LSST magnitude depth will discover so many new M Dwarfs, even a subset of potential stars will be massive. Defining and prioritizing appropriate targets in the LSST field will be necessary.
- The bimodality of M dwarf rotation periods means that many targets will likely be rapid rotators with $\sim$1 day periods. These may be challenging to identify in light curves from the first two years of LSST depending on the survey strategy due to the irregular observations and inter-night gap being greater than a day. This challenge will lessen with additional LSST data releases.
- Blending will make precise measurements difficult in crowded fields.

B.6.6.4. *Running on LSST Datasets (for the first 2 years)* —

- Starting with a smaller data subset would allow for trouble shooting of the survey strategy. One such data set could be drawn from MLSDSS-GaiaDR2 sample (Kiman

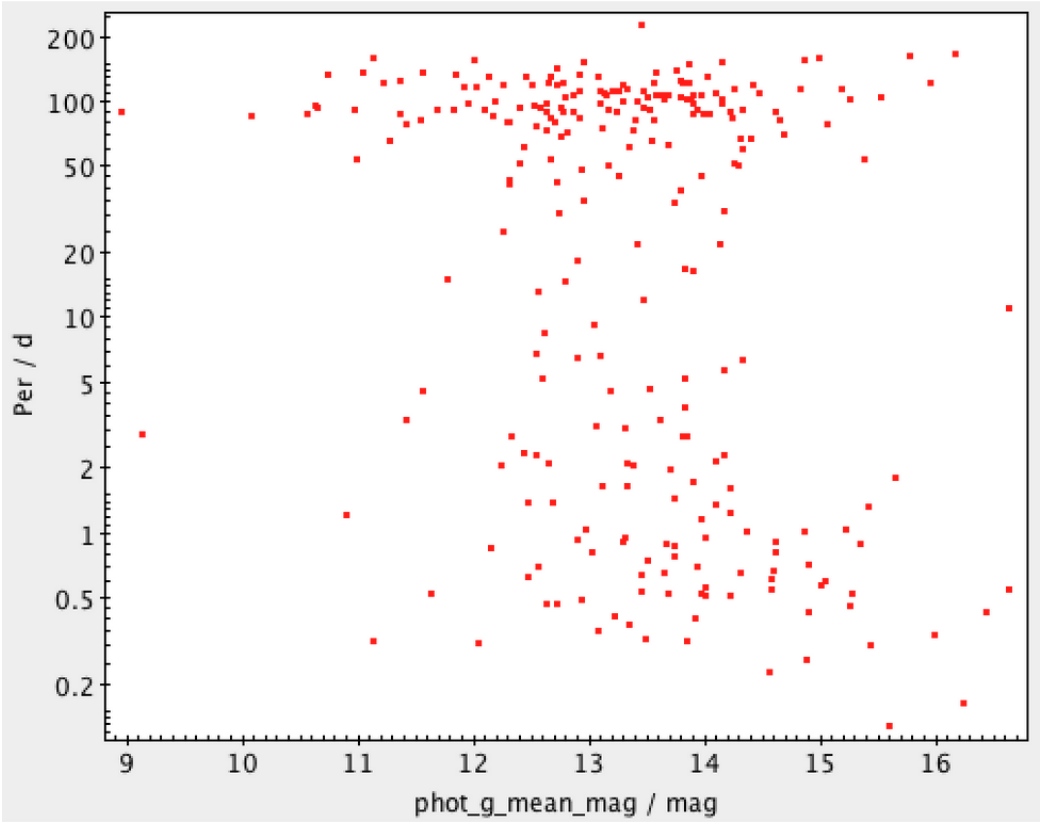

Figure 4. Period vs Gaia G mag for objects in the MEarth sample (Newton et al. 2018). Notice the bimodality of the rotation periods between ~1 and ~100 days. The brightness limit is based on the volume limit of the sample.

et al. 2019) which has < 20000 M Dwarf sources that have been matched with Gaia DR2 within the LSST field of view, most of which are fainter than Gaia G = 17 mag.

- The Wide Fast Deep survey will provide a sufficient cadence for identifying long rotation periods (> 100 d).
- Additional follow-up of suspected short period objects can speed up characterization.

B.6.6.5. *Precursor data sets* —

- The MEarth South data set provides a small sample (< 100) of known long rotation periods. While these sources will saturate in LSST observations, they provide an important reference for comparison with the LSST-discovered sample.
- Other large long-term surveys that overlap with the LSST field can be used to extend light curve coverage back and provide addtional characterization (e.g., ZTF, TESS, Pan-STARRS).
- MLSDSS-GaiaDR2 can serve as an initial sample.

B.6.6.6. *Analysis Workflow* —

- Identify and characterize M Dwarfs from color-color and color-magnitude diagrams. Older field M dwarfs are more likely to have longer rotation periods (e.g., Angus et al. 2020; Lu et al. 2021; Popinchalk et al. 2021), and so priority can be placed

on older objects that should appear fainter and bluer compared to the main sequence (see Kiman et al. 2019).

- Various periodograms will be run on each light curve to identify different types of variability. These will be re-run as new data points are added with each visit.
- Create phase-folded light curves determined from the best period based on each periodogram. Record amplitude of variability, most-probable period.
- A period-sensitivity analysis algorithm will need to be run to flag periods that are under explored due to irregular observation cadence.

B.6.6.7. *Software Capabilities Needed* —

- Interactive color-color and color-magnitude diagrams.
- Period detection algorithms that update as data points from subsequent visits are added.
- Interactive visualization of phase-folded light curves, with the capability to display user-supplied periods.
- An algorithm that flags periods that may be due to cadence windowing effects.

B.6.6.8. *References for Further Reading* —Newton et al. (2018), Kiman et al. (2019), Angus et al. (2020), Popinchalk et al. (2021)

B.6.7. *Identifying Substellar Companions to White Dwarfs*

Contributors: Alekzander Kosakowski

B.6.7.1. *Abstract*—Planets at distances of ~1 AU will be engulfed by, and potentially merge with, their host star during red giant evolution. More distant planets may be destroyed due to the effects of tides from their host star. However, massive planets and brown dwarfs may survive these events and end up on a compact orbit ($P \sim 1$ h) with an evolved star (see Rappaport et al. 2021). Additionally, more distant companions may avoid common-envelope evolution entirely, but still migrate to shorter periods ($P \sim 1$ d), potentially due to scattering interactions as the orbits of other planets become unstable during their host star's evolution.

Main sequence stars with masses $M \lesssim 8$ $M_\odot$ will evolve into white dwarfs, potentially leaving behind remnants of their planetary systems. In rare cases, these remnants can be observed as debris disks (Vanderbosch et al. 2020, 2021; Gentile Fusillo et al. 2021) or even as evidence for intact planetary bodies (Vanderburg et al. 2020; Blackman et al. 2021).

Planetary and sub-stellar companions to white dwarfs will cause rapid, deep eclipses in astronomical survey data with eclipse depths likely below the noise floor of each individual observation. Thus, the deep multi-band observations of LSST provide an ideal platform for identifying these deeply-eclipsing systems in the southern sky. Combined with ZTF in the north, and overlap with BlackGEM in the south, this project will provide insight to the fate of planetary systems around white dwarfs, potentially including formation rates of second-generation planets around evolved stars.

B.6.7.2. *Science Objectives*—This project aims to create a catalog of deeply-eclipsed white dwarfs with low-mass stellar and sub-stellar companions.

B.6.7.3. *Challenges (what makes it hard)*—

- White dwarfs eclipsed by sub-stellar companions are likely to show eclipses with depths below the noise-floor of a single observation, with eclipse depths approaching 100%.
  - Stacked images may provide estimates on the true mid-eclipse depth, but detecting enough mid-eclipse data points for sufficient image stacking is unlikely given the eclipse duration relative to the expected orbital periods.
- These systems are relatively rare; we can't simply ignore the Galactic plane and expect to obtain acceptable results. While the deep eclipses should still be visible in blended photometry, developing a proper deblending algorithm is still a large problem for this project.
- The expected eclipse duration of these systems is on the order of $1 \sim 10$ min, with orbital periods that may be longer than 1 day. Obtaining enough data to identify and constrain the periods of these systems through eclipses may require many months to years of data collection with LSST alone.

  - Combining data across many surveys will help mitigate this issue, but requires an efficient multi-survey cross-matching algorithm and support for importing data from non-public surveys and single-night user photometry.

- Obtaining follow-up spectroscopy for radial velocity measurements towards estimating companion masses requires a lot of telescope time, which may not be easy to obtain without a well-constrained orbital period from precise eclipsing light curve data.

B.6.7.4. *Running on LSST Datasets (for the first 2 years)*—While the LSST alerts system will provide valuable information towards identifying these deeply-eclipsing systems, many epochs will be required to determine orbital periods.

Until accurate periods can be determined, follow-up photometry observations will be unfeasible. Despite this, the validity of the LSST alerts can be easily verified by examining the fits images for a completely disappearing star. Follow-up spectroscopy will be completed on verified deeply-eclipsing white dwarfs. Systems which show infrared excess in their SED must contain low-mass stellar companions, providing a simple filter when targeting sub-stellar companions.

Because these systems can show nearly 100% eclipse depths, a sophisticated forced-photometry pipeline is important for this project. Photometry extraction with varying aperture sizes will be required for minimizing the noise level within the eclipses, specifically the ingress and egress where the flux is still measurable in a single exposure.

B.6.7.5. *Precursor data sets*—The ZTF coverage in the north, and especially its overlap with LSST in the south, provides a valuable prototype for this project. Methods applied to the existing ZTF data can be immediately applied to the LSST data.

While less deep in terms of limiting magnitude, BlackGEM will augment the LSST dataset, both in time-domain astronomy and mapping the southern-sky in multiple filters. Having access to both BlackGEM photometry and LSST photometry will significantly speed up the detection of the relatively bright objects detectable in both surveys.

B.6.7.6. *Analysis Workflow*—Measuring precise orbital periods from eclipsing systems requires many epochs of data. However, the LSST alerts combined with reference images for each field allows progress to be made on this project while waiting on sufficient data to be collected for orbital periods.

- First, a catalog of white dwarf candidates will need to be created using LSST colors. Gaia DR3 parallax can only be used down to $G \approx 20 \sim 21$ mag, but may still help create a color-region for selecting faint LSST white dwarfs.
- LSST alerts will be monitored closely for triggers from this LSST white dwarf catalog. Objects that show $\approx 100\%$ flux dips will be cataloged for potential follow-up observations.

- Follow-up spectroscopy will be obtained on objects without infrared excess in their SEDs. Companion masses will be estimated from radial velocity shifts measured in the white dwarf's spectrum.

After many months of operation, LSST will have enough data to perform a proper period search on these cataloged deeply-eclipsing white dwarfs. Analyzing the light curve data will roughly follow the same steps as other programs:

- Filter "poor-quality" images
- Deblend the photometry of affected objects in crowded fields.
- Perform forced photometry each white dwarf, using various aperture sizes to obtain the best signal-to-noise ratio per image.
- Run a modified box least squares periodogram on each system, searching for eclipses in candidates that have produced a certain number of alert triggers.
- Place limits on the true eclipse depth by creating co-added images using only mid-eclipse epochs.

B.6.7.7. *Software Capabilities Needed*—This project has similar software requirements as presented in section B.6.3 for identifying compact white dwarf binaries.

- This project will require a highly-optimized, modified box least squares algorithm to identify variability on both short ($P \sim 1$ h) and long ($P \gtrsim 1$ d) timescales. Thus, support for GPU acceleration is required to handle a wide and dense frequency grid.
- The ability to create user-defined sub-catalogs based on LSST alerts and photometry and astrometry from LSST and other sources, such as Gaia.
- A custom alert system capable of handling user-defined filters. Alerts for only objects present in user-defined catalogs will be sent to the user. In addition to basic alert information, alerts should record cumulative alert number and time since previous alert per object.

B.6.7.8. *References for Further Reading*—Blackman et al. (2021), Rappaport et al. (2021), van Roestel et al. (2021), Vanderbosch et al. (2020), Vanderbosch et al. (2021), Vanderburg et al. (2020)

B.6.8. *RR Lyrae Catalogs*

Contributors: Andy Connolly (ajc@astro.washington.edu)

B.6.8.1. *Abstract*—With a single visit depth of r ≈ 24.7 mag and ~1000 repeated observations over a 10-year period, LSST provides an opportunity to measure the distributions of RR Lyrae stars (RRL) within the Galactic disk and halo. The tight correlation between period, luminosity, and metallicity of RRL enables the calculation of accurate distances (~3%) out to >100 kpc within the first 2 years of LSST. This proposal is to create a catalog of RRL using the first 2 years of LSST data, to measure their metallicities and estimate the 3D metallicity distribution within the Galactic disk and halo. To accomplish this, RRL must be identified within the variable sources detected by LSST (either the alert stream or batch processing of the static data) and their periods (and any additional modulations to the light curves) measured. RRL have a typical period of 0.2–1.1 d and amplitudes of variation of $|sim$0.5–1 mag.

Discriminating between RRL subclasses will determine the accuracy of the period and metallicity estimates. RRL are divided into three subclasses: RRab which have a fundamental radial-mode pulsation with skewed, non-sinusoidal light curves with large amplitude; RRc which have a radial first overtone mode with sinusoidal variation but at a smaller amplitude than RRab; and RRd which simultaneously pulsate in both modes. Blazhko RRL show modulations in the light curve amplitude of ~0.1 mag over weeks to months. Blazhko RRL account for 20-30% of the total RRL population.

B.6.8.2. *Science Objectives*—

- Cataloging RRL from the LSST data and measuring their distances to $\lesssim$ 3% accuracy.
- Characterizing RRL distances and metallicities requires accurately measuring the periods of RRL from their light curves and identifying those RRL showing modulations due to the Blazhko effect.
- Estimating of the metallicity of the RRL from a Fourier analysis of the phased light curves.
- Identifying binary RRL within the catalog to estimate the masses of the pulsators.

B.6.8.3. *Challenges (what makes it hard)*—

- Identifying the variable sources will require iterative processing of time-series data to remove artifacts within the data and to improve the model/period fitting
- Data will continue to be updated as new observations of RRL are obtained.
- Running User Defined Function (UDF) at scale on LSST data (across all light curves)
- Sampling of RRL will be sparse, and the observations will be noisy. "For a distance modulus up to 19, for more than half of the survey footprint more than half of the light curves can be fit correctly using time intervals of 30 days. For a distance modulus of 21, we have to move to a time interval of 50 days to get a correct fit for 10%." Hernitschek & Stassun (2021)
- Period finding (multiband) can be slow

- Aliasing of periods from sparse sampling will be common in the early data
- Identification of those RRL affected by Blazhko modulation

B.6.8.4. *Running on LSST Datasets (for the first 2 years)* —

- We will analyze both the alert stream and the data release light curves. Most of the analysis will be based on catalog data but we will need to go back to postage stamp cutouts and potentially full fields to visualize any problems with the data.
- The initial analysis will process all variable point sources and then once RRL have been identified we will fit periods and metallicities to a subset of the data.
- We expect to detect $\sim 10^8$ variable stars in the first 2 years of Rubin with ~160 observations spread over $ugrizy$. The first analysis can be taken after 6 months of normal survey operations and will be reanalyzed every 3–6 months.
- We will use data from the Deep Drilling Fields to validate the analysis (due to its better sampling) and then the Wide Fast Deep data for the analysis.

B.6.8.5. *Precursor data sets* —Precursor data from Gaia and ZTF (bright but well sampled) and Pan-STARRS (multiband)

B.6.8.6. *Analysis Workflow* —Data cleaning and identification of artifacts within the data (this will be iterative as we progressively remove/flag bad data from the time-series catalogs):

- Remove poorly calibrated photometric data and sources flagged with suspicious photometry (e.g., source on edge of the CCD or a diffraction spike).
- Filtering undertaken using LSST DM source quality flags.
- Remove/flag outlier measurements from a light curve.
- Remove/flag extended sources.
- Visualize light curves and sequence of images to explore images and light curves for “bad" RRL.

Data storage and archives:

- Collect training samples and store in a DB (known RRL stars from Pan-STARRS, Gaia, ZTF)
- Collect training samples for metallicity measurements of known RRL
- Build representation of light curve (multiband) with time, passband, noise, flags for bad points from either the Level 1 database or Alert Stream
- Cross match to WISE (or other IR data) to improve RRL color selection

Variability Identification and RR Lyrae selection:

- Select sources with given number of epochs of data, and SNR (magnitude range)
- Characterize variability amplitude (e.g., Root-Mean-Square (RMS) of light curve) and filter based on this amplitude
- Measure a multiband structure function (Hernitschek et al. 2016)
- Separate RR Lyrae based on structure function (characteristic time, variability) and IR color (probably using a CNN approach). Expect 85% purity and 80% completeness

Period finding and metallicity estimation:

- Fit Lomb-Scargle multiband period finder and return periodogram and peaks in periodogram
- Apply prior to periods (0.2–1.1 d) to remove aliased periods
- For low SNR or poorly sampled light curves we can also fit RRL templates to estimate periods
- Estimate Fourier series for phased RRL light curves and fit metallicity measurements using ϕ_{31}

B.6.8.7. *Software Capabilities Needed* —

- Ability to apply selection filters to data (SQL query)
- Ability to run Lomb-Scargle fitter across all variable sources (e.g., as a UDF using Apache Spark)
- Storage of light curves as objects (time, passband, noise, flags for bad points) with annotations (e.g., classifications) that can be queried and cross matched to other data sets
- Storage of outputs of filtered and classified data (or flags based on filters applied to existing catalogs)
- Visualization of distribution of properties of sources (e.g., color-color scatter plots and histograms) colored by flags
- Visualization of distributions of selected sources on the sky and relative to camera coordinates
- Visualization of postage stamps and light curves for individual sources

B.6.8.8. *References for Further Reading* —Sesar et al. (2017), Hernitschek et al. (2016)

B.6.9. *Exceptional Variability: New Astrophysics & Technosignatures*

Contributors: James R. A. Davenport (jrad@uw.edu)

B.6.9.1. *Abstract*—Rubin provides an unmatched ability to carry out searches for truly exceptional forms of stellar variability, which are sure to challenge our understanding of stellar evolution. These may take the form of e.g. dramatic outbursts or dimming from stars that were thought to be "non-variable", subtle changes in stellar properties over the LSST baseline, or unexplained variability on a variety of timescales (e.g., "Boyajian's Star", Boyajian et al. 2016). These behaviors may not be remarkable as compared to the vast array of light curve morphologies that Rubin will observe for stars, but when placed in context (e.g. compared to similar stars on the H-R Diagram) they would reveal themselves as outliers.

Perhaps the most extreme form of exceptional variability would come from a technosignature – a signal or byproduct of extraterrestrial technological activity. These signals may be subtle in amplitude, but highly unusual in timing, for example. The most unusual behavior from otherwise "boring" stars must be scrutinized under this lens, which will motivate extensive follow-up study.

B.6.9.2. *Science Objectives*—

- Identifying new forms of stellar variability on all timescales
- Identifying exceptional or unique star systems compared to other similar stars
- Making robust parameter space constraints from technosignature searches with Rubin data

B.6.9.3. *Challenges (what makes it hard)*—

- Outlier detection in the behavior of stars. Sparse sampling makes the distribution of possible "normal" behavior for stars very large.
- Developing scalable, appropriate technosignature algorithms. Identifying those from the e.g. Radio astronomy community that are applicable to optical surveys.
- Outlier detection with < 100 epochs is challenging.

B.6.9.4. *Running on LSST Datasets (for the first 2 years)*—

- The alert stream will be useful for monitoring stars along the "SETI Ellipsoi" (Lemarchand 1994) or other favorable directions for unusual behavior, e.g. the Earth Transit Zone (Heller & Pudritz 2016).
- The deep drilling fields will provide the best empirical models of what "normal" behavior for stars is.
- First challenge is to find exemplars, things that stand out as being patently unusual with < 100 epochs.
- Final challenge is to "Classify Everything", and whatever remains is therefore exceptional.

B.6.9.5. *Precursor data sets* —

- ZTF possibly the best precursor dataset given the sampling and baseline, but many systematics to consider in identifying stellar outliers.
- Having a robust set of templates for every "type" of star would be ideal, akin to the "One of Everything" (Lacki et al. 2021) catalog. This could be drawn from high-cadence data from e.g. TESS (Ricker et al. 2015).

B.6.9.6. *Analysis Workflow* —

1. Gather large sample of stars, estimate rough stellar parameters, place in context on the color–magnitude diagram (CMD)
    - Better yet: use the color-color-color-color-magnitude diagram, and include upper limits for colors at the extreme red or blue end.
    - include extinction corrections and their uncertainties in this space.
2. Dynamically segment the CMD into small regions of self-similar stars
3. Compute a large number of features for every light curve, including variability metrics, Lomb-Scargle timescales, etc.
4. Actively hunt for outliers in each CMD bin. What are the most unusual stars relative to their siblings?
    - We are building towards the goal of, for ~10B stars, being able to immediately say: What kind of star is this? What is the limit of "normal" behavior for a star like this?
5. Coordinate alert monitoring with other targeting approaches (e.g. technosignature methods)

B.6.9.7. *Software Capabilities Needed* —

- Ability to create a reliable Color-Magnitude Diagram, including extinction corrections, for ~10B stars without precise distances from e.g. Gaia.
- Ability to quickly cross-match & place a star from e.g. the Alert Stream into a CMD bin
- Ability to generate multi-band variability features (e.g. general light curve stats, Lomb-Scargle periods, fast Gaussian Process predictions)
- Ability to cross match catalogs at scale, for alerts and known stars
- Ability to quickly compute 2-D and 3-D separation between objects across the sky
- Ability to do feature selection, outlier detection, and basic machine learning at scale

B.6.9.8. *References for Further Reading* —SETI Ellipsoid: Lemarchand (1994)
Earth Transit Zone: Heller & Pudritz (2016)
SETI w/ surveys: Djorgovski (2000)
"Boyajian's Star": Boyajian et al. (2016)
"One of Everything": Lacki et al. (2021)

Systematic Serendipity: Giles & Walkowicz (2019)
SETI in the Spatio-Temporal Survey Domain: Davenport (2019)

B.7. *Solar system science*

B.7.1. *Non-Tracklet Discovery for Small Body Populations*

Contributors: Joachim Moeyens (moeyensj@uw.edu)

B.7.1.1. *Abstract* —The Vera C. Rubin Observatory's Legacy of Survey of Space and Time (LSST) will discover over 5 million new Solar System small bodies over the course of its 10-year survey. The Solar System pipelines that will enable these discoveries will rely on observing "tracklets": two-dimensional sky-plane motion vectors that constrain the position and rate of motion of moving objects (Kubica et al. 2007; Denneau et al. 2013; Jones et al. 2018b; Holman et al. 2018). A tracklet requires at least two observations to be made in a single night typically within 90 minutes. Three such tracklets are required over the course of a 15-day linking window to identify the presence of a moving object. The requirement to observe tracklets has two immediate consequences: 1) to successfully discover minor planets tracklets must be observed which imposes a strong constraint on cadence, 2) any minor planets that are not observed in at least three tracklets over a 15-day window will not be discovered by Rubin Observatory pipelines. For example, Near-Earth Object (NEO)s may move too quickly for a tracklet to be formed, and at the extreme end, very distant objects may not exhibit sufficient discernible motion within 90 minutes for two observations to be identified as separate. Tracklet-less Heliocentric Orbit Recovery (Moeyens et al. 2021) is a small body discovery algorithm capable of discovering Solar System small bodies without the need to use tracklets. The algorithm currently focuses on discovering Main Belt asteroids and trans-Neptunian objects, with future extensions planned to tackle the NEO population. While capable of discovering minor planets without tracklets, THOR requires significantly more computational power than traditional tracklet-based algorithms. The Asteroid Institute, a program of the B612 Foundation, has started work to deliver the Tracklet-less Heliocentric Orbit Recovery, an algorithm described in Moeyens et al. (2021) (THOR) algorithm as part of their Asteroid Discovery, Analysis, and Mapping (ADAM). ADAM is a scalable, cloud-based astrodynamics platform designed to enable large-scale analyses with small body science in mind. The aim of the discovery service is to provide the community with a tool to search for asteroids in any astronomical dataset, and ultimately, to find the asteroids that may have been missed in the LSST dataset.

B.7.1.2. *Science Objectives* —

- Filter the LSST alert stream to contain only unattributed observations. Additionally, filter out false positive detections using tools such as a real-bogus filter
- Create an automated test orbit selection algorithm to maximize discovery space. This selection algorithm should scale with data properties such as density of observations and the phase space density of observed small body populations
- Run discovery pipeline and validate results
- Submit discovery candidate observations and observations of known objects to the Minor Planet Center (MPC)

B.7.1.3. *Challenges (what makes it hard)* —

- LSST's depth will push the computational power needed to successfully identify small bodies
- Collaboration/communication between the discovery service and the internal Rubin pipelines: we do not want to compete for discoveries but instead we want to deliver complimentary analyses so that we maximize discovery potential. There are a number of questions in defining this interaction. Do we ignore tracklets in the alert stream? Is there a way to get access to the observations that were combined into tracklets? Do we include a tracklet-builder as part of the discovery service?
- The false-positive density in the alert stream (and in difference images) may have a significant impact on the processing time, particularly orbit determination.
- Depending on the linking window size (15 days or more), processing should occur nightly as new observations are made. This would require being able to query for the current night's observations and the previous 14 nights of observations quickly.
- Longer baseline discovery searches should be encouraged to maximize the discovery potential for very distant objects (effectively increasing the linking window to ~30 days or longer), doing so will add additional processing time and the need for querying for larger volumes of observations.
- To avoid pipeline bottlenecks, the discovery algorithm hosted in the cloud should return results before the next night's observations.
- THOR needs to be extended to the NEO population since initial development focused on easier-to-link populations. The on-sky motion of NEOs may increase the computation cost of THOR by an additional factor of 10.
- Orbit determination techniques can decrease the computational cost but still need development: the current technique falls back to on-sky coordinates and ignores much of the benefits gained by linearizing the linking problem relative to the motion of test orbits.

B.7.1.4. *Running on LSST Datasets (for the first 2 years)* —

- LSST Alert Stream (or difference image sources) initially
- Possibly the LSST observation catalogs for longer-baseline discovery searches

B.7.1.5. *Precursor data sets* —Current datasets used for development:

- THOR was initially developed and tested on two weeks of ZTF observations (specifically, the alert stream hosted at University of Washington (UW))
- The discovery service is being actively developed with the goal of processing the NOIRLab Source Catalog (DR2) (Nidever et al. 2021)

Future datasets to search and use for development:

- Pan-STARRS is a possible dataset to mine for more discoveries
- The simulated LSST dataset is a good candidate for testing to LSST scale

B.7.1.6. *Analysis Workflow* —

- Nightly Operations
  - Trigger discovery search on new nightly observations
  - Gather previous 15 nights' worth of observations
  - Filter alert stream to discard non-moving object sources and observations of known moving objects, filter out as many false positive observations as possible
  - Calculate optimal test orbits and submit discovery job to cloud-based infrastructure
  - In a test scenario, observations of known moving objects should not be discarded so that algorithmic completeness can be calculated
- >Monthly Operations
  - Trigger discovery search on monthly threshold of observations
  - Gather previous 30 or more nights' worth of observations
  - Query the LSST observations catalog for unassociated observations that span ~months, apply filters to remove non-moving object sources and observations of known moving objects
  - Calculate optimal test orbits and submit discovery job to cloud-based infrastructure
- Discovery search launches in the cloud (for both nightly and monthly operating modes) and produces a list of candidate discoveries (their constituent observations and orbits) and observations of known objects that may not have been correctly identified as such
- Discovery candidate observations and observations of known objects are submitted to the MPC

B.7.1.7. *Software Capabilities Needed* —

- Fast/reliable querying of the LSST alert stream
- Fast/reliable querying of the LSST observation catalogs on monthly cadences
- Filters to remove static and known sources from the LSST alert stream, tools to filter out or minimize the presence false positive detections in the alert stream
- Identification of tracklets in the alert stream (if LSST tracklet building can be accomplished near real-time)
- THOR discovery performance extended to the NEO population with enhancements for orbit determination and test orbit selection
- Continued development of cloud-based infrastructure underlying ADAM
- Task-queue system with autoscaler for THOR discovery jobs (parallelized by test orbit or chunks of sky or both)
- Visualization tools to track in-progress discovery searches

- Speed enhancements in THOR and cloud-based infrastructure to reduce computational cost with cost-benefit analysis of cores/machines vs processing time to handle LSST data volume

B.7.2. *Characterizing Populations of Active Small Bodies*

Contributors: Orion Chandler (orion@nau.edu), Henry H. Hsieh (hhsieh@psi.edu), Agata Rożek (a.rozek@ed.ac.uk)

B.7.2.1. *Abstract*—Small solar system objects can exhibit activity, or visible mass loss, due to a variety of mechanisms. The most common of those mechanisms is the sublimation of volatile ices, which drives the activity of the vast majority of known comets. While typically associated with comets from the outer solar system, sublimation-driven activity has also been observed on a small number of objects in the main asteroid belt (known as main-belt comets), presenting intriguing new opportunities for studying the origin of terrestrial water. Recently, comet-like activity attributed to mechanisms such as impacts or rotational destabilization has also been detected. Activity produced by these mechanisms is unpredictable and transient, lasting anywhere from several months to just a few days. If identified promptly, these active events present rare opportunities to perform observational studies of processes that are largely only studied using theoretical or computational models, or in laboratory settings that can only approximate certain aspects of those processes in the natural world.

With its unprecedented imaging sensitivity and sky coverage, LSST has the potential to revolutionize active solar system object science by detecting activity that has been too weak or too short-lived to be reliably detected by other surveys, and providing regular deep monitoring of known active objects that cannot currently be done for large numbers of objects. LSST will be able to discover cometary activity at much larger distances than current surveys, enabling long-term studies of incoming dynamically new comets as they pass through different regions of the solar system. It should also greatly increase the number of impact- or rotationally-driven active events that are detected in time to conduct real-time observations, increase the number of known members of populations of rare objects like main-belt comets and active Centaurs, and potentially uncover active objects in small-body populations in which no active objects are currently known (e.g., the Jupiter Trojans).

Achieving these myriad goals, however, will require LSST to have automated activity detection algorithms that can promptly search all moving objects detected in a night for wide range of activity morphologies, and trigger observational follow-up for confirmation or detailed activity characterization. Other challenges associated with activity detection and characterization include the potential need for larger image cutouts than are currently planned to be provided as part of LSST alert packets for all solar system object detections, and the need for a system that can track the outcomes of non-pipeline analyses (e.g., human vetting or follow-up observation results) and link them to the appropriate solar system objects in the LSST or broker databases.

B.7.2.2. *Science Objectives*—Active object science with LSST consists of two major components: the discovery of new active objects and the characterization and long-term monitoring of known active objects. LSST has the potential to revolutionize active object science in

both of these areas, but substantial software challenges will need to be met for the survey to achieve this potential.

The LSST will discover a large number of active bodies from myriad known populations, including active asteroids, active Centaurs, and active NEOs. Given its unprecedented sensitivity, the survey may also discover activity in other small body populations that are not currently known to contain active bodies, such as the Jupiter Trojans and trans-Neptunian objects (TNOs). LSST will be especially well-positioned to identify long-period comets and interstellar comets at large heliocentric distances, which will enable study of dynamically new bodies as they enter the inner solar system for the first time, and identify candidate targets for the European Space Agency (ESA) Comet Interceptor mission. By exploring the dynamical and physical parameter spaces of all known and newly found active bodies, we will be able to better understand the conditions required to produce activity. To achieve that goal, we will quantify activity occurrence rates in various populations as functions of heliocentric distance and orbital position, and also take into account other factors such as activity strength, recurrence, and dust dynamics.

For all objects, we will measure activity properties (e.g., dust production rates, tail morphologies) and track these properties as functions of time to enable diagnosis and understanding of underlying activity mechanisms. We will also conduct forward and backward dynamical integrations to trace the origins and future dynamical evolution of active bodies. These dynamical insights will give context to the presence of active objects within each dynamical population and inform sublimation modeling (see Chandler et al. 2020) that can enable estimates of which molecules are most likely responsible for observed sublimation-driven activity.

These objectives should be achievable by LSST if software can be successfully deployed that detects activity (or any form of mass loss) for small solar system bodies over a wide range of heliocentric distances, brightnesses, morphologies, and time periods, and is also able to conduct uniformly-defined quantitative analyses on known active objects with extended morphologies (e.g., for detecting events like cometary outbursts, which are characterized by rapid increases in brightnesses of cometary comae). Current surveys are limited by relatively shallow single-visit image depths in addition to the inherent challenges of activity detection in survey data and uniform analysis of extended objects that can vary widely in morphology. At present, identifying activity requires a mix of automated flagging procedures and human vetting. The LSST single-visit image depths will surpass those of other ground-based surveys, yet significant challenges with automated flagging and human vetting will remain, and these may be exacerbated by the high-volume data flow LSST is expected to produce.

B.7.2.3. *Challenges (what makes it hard)*—The full suite of activity detection software to be used for LSST data analysis will ideally include a variety of automated activity detection techniques in order to account for a range of activity morphologies. Examples of different morphologies we would like to be able to detect include coma strong enough to affect an object's PSF Full Width at Half-Maximum (FWHM), low-level activity too

faint to affect an object's PSF FWHM, completely unresolved activity that only affects an object's photometry, circularly symmetric activity, and asymmetric or highly directed activity. Automated tools exist for flagging the potential presence of some of these types of activity morphologies, but not all, and human vetting is still required to confirm the presence of activity in essentially all cases. For LSST, new software will need to be developed to automate searches for activity using specific approaches that currently require some degree of human intervention, while other code will need to be enhanced to work at LSST scales, both in terms of data processing speed and method of ranking and prioritizing results for any needed follow-up.

Software will also need to be developed to perform uniformly-defined quantitative analyses of known active objects in such a way that derived quantities can be used to quantify an active object's evolution over time and make comparisons to other active objects. Such characterization tasks can be relatively simple, such as measuring the amount of flux within a radius of a certain angular or physical distance from the nucleus, or substantially more complex, such as characterizing the morphology of an active object's comae (e.g., identifying whether one or more tails are present or not, and potentially quantifying the lengths, directions, and shapes of those tails).

In terms of scale, all activity detection algorithms will need to be applied to all solar system object detections made each night (expected to be on the order of 1 million). This approach is needed to avoid selection biases, and is also necessary for many activity detection algorithms themselves due to their need to establish templates for inactive moving objects to which potentially active outliers can be compared. In contrast, however, advanced activity charaterization tasks will only need to be applied to detections of objects that are already known to be active, which should comprise a far smaller data set.

Further increasing the computational demands of active object analysis, many activity detection and characterization algorithms will require access to image data (see Appendix B11, pg 30, of Hsieh et al. 2019). Notably, some algorithms may require larger image cutouts than will be available from the LSST alert stream, though exact minimum viable cutout sizes for each algorithm are yet to be determined. Crucially, it will be essential to process the image data rapidly in order to keep pace with image acquisition and to enable follow-up observations for confirmation and characterization.

Human vetting and even follow-up observations will be required in many cases to confirm activity detection results from LSST, especially early in the survey while the application of detection algorithms to LSST data and the algorithms themselves are being refined based on their early performance. This adds subjectivity and external elements to pipeline-centered search efforts. A mechanism for keeping systematic records of any human vetting and follow-up observation results will need to be developed in order to facilitate on-the-fly algorithm evaluation and improvements, as well as later debiasing efforts.

Mechanisms for taking past activity scores of objects into account when searching for activity would be very useful to have if possible, given that multiple recent marginal detections of activity can collectively comprise a stronger detection of activity than any

of those detections alone. This will require activity detection code to have the ability to link past data for a given object to current data, and incorporate those data into activity evaluation procedures.

We anticipate eventually making use of machine learning techniques to improve activity detection algorithms, and potentially remove the need for human vetting, but these techniques are not currently in widespread use for this purpose. Additionally, some ML techniques, such as neural networks, require large labelled training data sets for training and testing. Consequently, software and a training data set will need to be developed for this purpose, further underscoring the need for careful tracking of all activity detection-related actions and decisions, automated or otherwise.

The ultimate quality of LSST difference imaging is currently uncertain, but it is already expected that it will not be particularly good early in the survey when sufficient data for developing high-quality sky templates for static sources have not yet been obtained. As such, given the very high probability of blending of solar system objects with background sources (due to LSST's unprecedented image depth), procedures will need to be developed to distinguish spurious activity detections from image artifacts.

B.7.2.4. *Running on LSST Datasets (for the first 2 years)*—Alert stream data, particularly "postage stamp" image cutouts and photometry, will be used to identify and (to some extent) characterize active object candidates and their activity. However, additional custom cutout image data may be required for specific activity detection and characterization algorithms and techniques.

Most image-based activity detection algorithms will, in principle, be able to run on survey data on Day 1 (if the software is ready) as they primarily use other sources in the same field for comparison in order to identify active objects. It should be possible to conduct many tasks related to characterization of known active objects from Day 1 as well, since they generally only require photometric analysis of single-visit images as long as the positions of those objects in the images can be identified. The lack of a complete set of template images for difference imaging analysis early in the survey is expected to limit transient detection during this period, and will present a challenge from Day 1 of the survey. Nonetheless, some mitigation can be achieved by targeting known solar system objects with exceptionally small ephemeris uncertainties until Difference Image Analysis (DIA) is fully operational. Activity characterization efforts may also be hindered if nearby or blended background sources affect photometry measurements (especially for objects that are particularly extended), but some mitigation of this problem can be achieved by simply ignoring images where objects are close to known background sources or are in high-density star fields in general, or by adding functionality to activity characterization software to flag outlying data that could be due to background contamination.

Photometric searches for activity will require temporal baselines of adequate duration in order for the anticipated solar system object photometric behavior to occur, thereby enabling outlier detection. Thus this process will be less efficient early on in the survey, although

such searches may still be possible using photometric data obtained for known asteroids by other past and concurrent surveys.

Some LSST data products will be useful for this work, such as PSF moments for PSF comparisons and photometry and phase functions for photometric detection of activity. Nonethelesss, significant development of new software tools for producing additional data products will be required to address, for example, other activity morphologies like unusual tails, or very faint and irregular comae. Also needed is infrastructure to achieve broader science objectives. Examples include human vetting in a highly systematic and documented fashion, prioritizing and triggering observational follow-up at facilities temporally and geographically appropriate given target orbital positions.

While clear activity may be detectable in a single image, multiple images will be required to satisfy the MPC's comet classification requirements. As of this writing, the MPC policy states that unnumbered objects require multiple nights with multiple images acquired within 1-2 days.

B.7.2.5. *Precursor data sets* —At present there is no single data set or service comparable to what the LSST and its brokers will supply. Data sets with comparable image depths include public archives of DECam data (hosted at National Science Foundation (NSF)'s NSF's National Optical-Infrared Astronomy Research Laboratory; https://nationalastro.org (NOIRLab) AstroArchive[5]), MegaPrime data (hosted at the Canadian Astronomy Data Centre (CADC) archives[6]), and SuprimeCam and HSC data (if the data are reduced). The ZTF Alert stream provides a comparable alert stream data product that includes preliminary analyses (e.g., extendedness) as well as template subtracted data. An alert broker simulating time domain events at similar depth to the LSST can be constructed to utilize publicly available DECam data. The data selected would necessarily have been acquired over a period greater than one month, such as DES and/or The Dark Energy Camera Legacy Survey (DECaLS). Extending Broker analyses to include PSF and extendedness would provide a similar service as to what is needed from the LSST data stream for this science case. A vetted dataset of active objects within this dataset would be needed to adequately test activity detection and characterization techniques.

B.7.2.6. *Analysis Workflow* —

1. Identify known and new moving object in nightly data
2. Retrieve sufficient surrounding image data as required for activity detection and characterization tools
3. Apply various activity detection algorithms (e.g., photometric enhancement analysis, PSF analysis, multi-aperture photometry analysis, wedge photometry, NoiseChisel; see Figures 5 and 6). As all of these analyses will be performed on individual detections, this work should be highly parallelizable.

[5] https://astroarchive.noirlab.edu
[6] https://www.cadc-ccda.hia-iha.nrc-cnrc.gc.ca

4. Apply various activity characterization algorithms (e.g., $Af\rho$ calculations, surface brightness profile characterization, automated basic dust modeling analysis, temporal image subtraction for outburst detection) and identify any unusual changes in those parameters that could indicate the onset of events such as cometary outbursts. As all of these analyses will be performed on individual detections, this work should be highly parallelizable.
5. Compute combined weighted activity confidence metrics for new active object candidates
6. Retrieve activity confidence metrics for previous detections of new active object candidates and increase priority for objects considered to be active candidates for multiple recent detections, even if activity metrics for individual detections have low confidence levels
7. Prioritize activity candidates for additional investigation to confirm and characterize activity via:
   (a) Observational follow-up
   (b) Searches for and analysis of archival data
   (c) Citizen Science classification and activity vetting
8. Prioritize cometary outburst candidates for observational follow-up to confirm outburst detections and analyze confirmed outburst events
9. Record results of confirmation analyses for new activity candidates and cometary outburst candidates, and link those results to the appropriate detections in LSST or broker databases
10. Incorporate results from initial activity detection analyses and confirmation analyses into ML-based tools
11. More advanced activity detection in the future might involve shifting (and possibly rotating) and stacking of multiple detections of the same object, or even of multiple objects (e.g., that share certain physical or dynamical characteristics, such as belonging to the same collisional family), to search for ultra-faint activity, so infrastructure for facilitating this would be desirable, but this is considered lower priority at the moment relative to single-visit activity detection and characterization efforts.

Our overarching goals will be to make use of various activity detection tools on all LSST solar system object detections and return parameters related to the activity confidence levels of those detections for each algorithm, and apply activity characterization tools on known active object detections to quantitatively characterize activity evolution to enable both quantitative studies of long-term activity evolution and identification of unusual changes in that evolution (e.g., outbursts). Activity and outburst detection parameters may include detection-level confidence parameters and object-level confidence parameters (e.g., parameters indicating whether an object consistently shows indications of activity or an outburst over multiple recent detections). Activity or outburst likelihood parameters should be automatically associated with their corresponding solar system objects in LSST or bro-

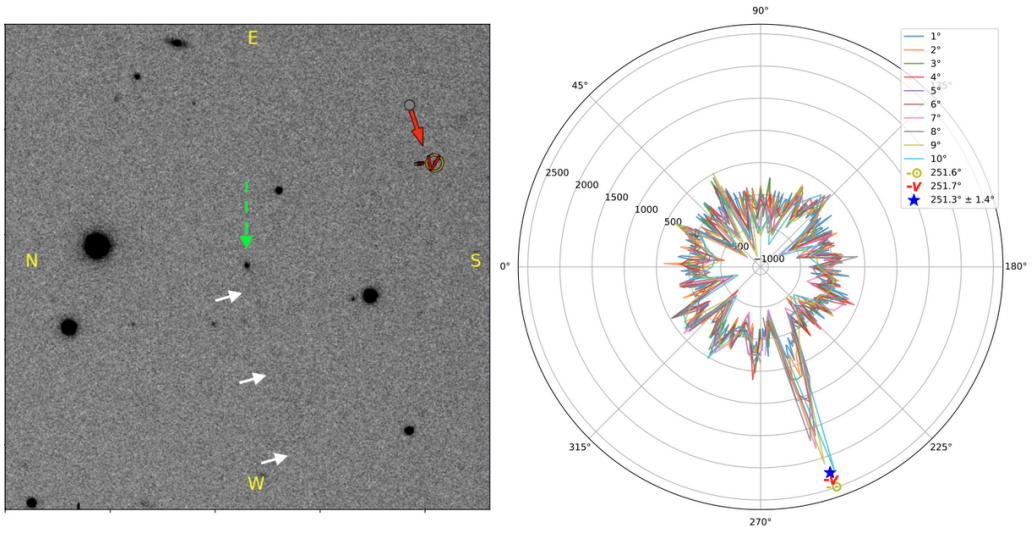

Figure 5. Left: active asteroid (248370) 2005 QN_{173} (green dashed arrow) with a tail (white arrows). Right: the application of a "wedge photometry" tool designed to detect tails by measuring counts within variable width bins (wedges of an annulus). The wedge measurement of the tail angle is in close agreement with the anti-Solar and anti-motion vectors computed by JPL Horizons. From Chandler et al. (2021).

ker databases, while results of any additional follow-up analyses to confirm the presence of activity or an outburst should be recorded and also linked to the appropriate objects in available databases.

B.7.2.7. *Software Capabilities Needed*—An alert broker will be needed for data integration, such as object matching with external catalog data, photometry, and activity likelihood parameters. We may make use of training sets developed using vetting by professional astronomers, follow-up observations, and Citizen Science projects (e.g., *Active Asteroids*[7]) to train ML-based image pattern recognition algorithms. We expect the computational overhead to be negligible for broker-provided analyses, however the computational cost for additional analysis tools, especially ML-based approaches, is unknown.

Image cutouts of each activity candidate will need to be stored for myriad reasons, including adherence to applicable MPC reporting mandates. Storage requirements have yet to be determined.

We do not anticipate the need for custom visualization tools. Widely adopted visualization tools like Matplotlib[8] and Bokeh[9] should suffice for static and dynamic/interactive data visualization, respectively.

Software implementations of some activity search algorithms do exist, such as those outlined in Appendix Byte (8 bit) (B).11 of Hsieh et al. (2019). However, most need to be optimized to run at LSST scale and some will need to be developed from scratch. To our knowledge, no software yet exists to perform the type of logging that will be needed

[7] http://activeasteroids.net
[8] https://matplotlib.org
[9] https://bokeh.org

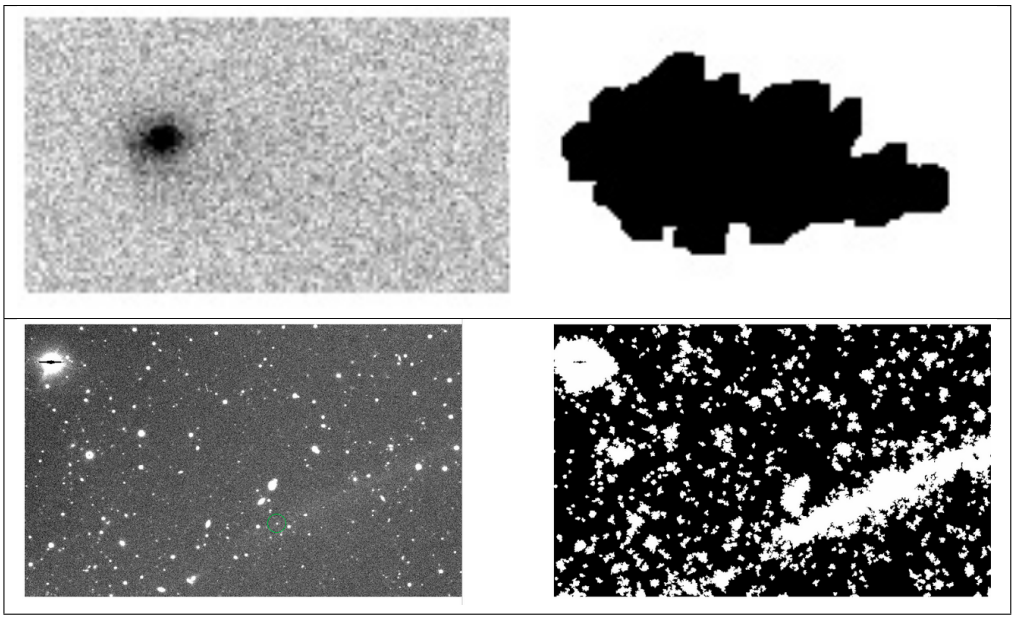

Figure 6. Example results from the tool NoiseChisel (Akhlaghi & Ichikawa 2015; Akhlaghi 2019). Top: Comet 358P/Pan-STARRS with a tail and coma that are difficult to identify in the original image (left), but the resulting NoiseChisel output (right) provides clear evidence of activity. Image courtesy Mohammad Akhlaghi and Henry Hsieh. Bottom: The original image (left) of comet 67P/Churyumov–Gerasimenko (green circle) with a faint tail extending roughly ENE. The NoiseChisel output (right) provides much more contrast, important for manual analysis by humans and automated searches by computers. Image courtesy Agata Rożek.

to record the outcomes of human vetting and follow-up observations to confirm activity detections for the purposes of later debiasing analyses, especially at the scale of the LSST survey.

We emphasize that cutout images larger than the postage stamps that will be automatically generated for all transient detections will be required for some activity detection and characterization analyses, meaning that custom extraction of image data, potentially for all solar system object detections every night, will be needed.

B.7.2.8. *References for Further Reading* —**Science Background:**

Chandler et al. (2020), "Cometary Activity Discovered on a Distant Centaur: A Nonaqueous Sublimation Mechanism", ApJL, 892, L38

Hsieh & Jewitt (2006), "A Population of Comets in the Main Asteroid Belt", Science, 312, 561

Jewitt et al. (2015), "The Active Asteroids", Asteroids IV, 221-241

Snodgrass et al. (2017), "The Main Belt Comets and Ice in the Solar System", Astron. Astrophys. Rev., 25, 5

Software/Algorithm Background:

Akhlaghi & Ichikawa (2015), "Noise Based Detection and Segmentation of Nebulous Objects", ApJ, 220, 1

Akhlaghi (2019), "Carving out the low surface brightness universe with NoiseChisel", The

Realm of the Low-Surface-Brightness Universe, Proc. International Astronomical Union (IAU) Symposium No. 355
Chandler et al. (2018), "SAFARI: Searching Asteroids for Activity Revealing Indicators", Publications of the Astronomical Society of the Pacific (PASP), 130, 114502
Chandler et al. (2021), "Recurrent Activity from Active Asteroid (248370) 2005 QN173: A Main-belt Comet", ApJL, 922, L8
Gilbert & Wiegert (2009), "Searching for Main-belt Comets Using the Canada-France-Hawaii Telescope Legacy Survey", Icarus, 201, 714
Hsieh (2009), "The Hawaii Trails Project: Comet-hunting in the Main Asteroid Belt", A&A, 505, 1297-1310
Hsieh (2015), "The Main-belt Comets: The Pan-STARRS1 Perspective", Icarus, 248, 289-312
Hsieh et al. (2019), "Maximizing LSST Solar System Science: Approaches, Software Tools, and Infrastructure Needs", arXiv:1906.11346
Sonnett et al. (2011), "Limits on the Size and Orbit Distribution of Main Belt Comets", Icarus, 215, 534
Waszczak et al. (2013), "Main-belt Comets in the Palomar Transient Factory survey - I. The Search for Extendedness", Monthly Notices of the Royal Astronomical Society (MNRAS), 433, 3115

B.7.3. *Constraining the Number Density and Mass of the Galactic Interstellar Small Body Reservoir*

Contributors: W. Garrett Levine (garrett.levine@yale.edu)

B.7.3.1. *Abstract*—We estimate the mass of small bodies per unit stellar mass that is ejected from planetary systems into the interstellar medium as rogue objects. In particular, we combine the first two years of Rubin/LSST data with the results from n-body simulations of exoplanetary systems. This initial LSST dataset is especially helpful because the expanded field-of-view means that the Interstellar Object (ISO) detection rate should be highest during the initial survey stages. Estimating the reservoir of ISOs is a question of detectability and requires a rigorous debiasing of the population of interlopers traversing the Solar System discovered by LSST. Although this calculation is most dependent on orbital parameters and asteroid size, small body shape and composition could be concerns as well. For this science case, we leverage the Rubin Science Platform and real LSST images with synthetic injected sources to construct forward models of ISO detections by the survey and fit these to the results of our orbital simulations. Through this study, we estimate the prevalence of dynamical instabilities like that of the Nice Model and put preliminary constraints on the occurrence and multiplicity of giant planets with Safronov numbers larger than unity. Because this calculation is based solely on ISO occurrence rates, this work probes the architectures of extrasolar systems independently from either dedicated searches for exoplanets or observations of circumstellar disks.

B.7.3.2. *Science Objectives*—The purpose of this science case is to develop an estimate of the Galaxy-wide aggregate interstellar small body ejecta. This objective can be broadly divided into two steps: modeling the ISO detection efficiency of Rubin/LSST, and conducting numerical simulations of dynamical instabilities in exoplanetary systems. The first task will result in a robust estimate of the ISO number density with well-constrained uncertainty, while the second one will couple observational results with theoretical astrophysics to interpret the LSST data in the context of planetary formation. These numerical simulations could constrain the size, composition, number, and age distributions of ISOs through forward modeling with the Solar System science collaboration's post processing pipeline to model detections.

The nature and composition of ISOs will be constrained through concentrated follow-up efforts; this question is outside of the scope of this study aside from any detectability effects on the types of ISOs which will be identified. For example, the typical Galactic kinematics of ISOs will depend on their compositions and characteristic ages.

B.7.3.3. *Challenges (what makes it hard)*—Many of the fundamental challenges for this topic are more generally addressed by other science use cases. For example, constructing comprehensive selection functions to quantify the small body detection efficiency will be critical to debiasing the census of LSST detections. However, the parameter space for hyperbolic orbits is significantly larger than for the set of elliptical orbits that are bound to the Sun. Additional selection functions may need to be computed to estimate the occurence

and size distribution of interstellar ejecta. In addition, objects on hyperbolic orbits are subject to substantial apparent motion in the sky; this effect could tangibly modify the small body detection efficiency for constant phase function and brightness.

The known population of ISOs currently consists of only two members. ‘Oumuamua and Borisov were markedly different objects with starkly contrasting orbital properties, so these objects could represent a bimodal distribution of ISOs, two points on a spectrum, or even outliers. Since LSST may expand this catalog by an order-of-magnitude, this science case also encompasses inherent “unknown unknowns.” ISOs will be only a small fraction of LSST’s total small body detections, so each identification will be critical to constraining the galactic number density. Therefore, identifying faint objects is especially important for this science case.

Finally, the nature of ‘Oumuamua-like ISOs is unknown. The Pan-STARRS survey showed that these objects may comprise a substantial fraction of interstellar small bodies, so LSST pipelines must be capable of efficiently detecting objects with highly variable lightcurves (via shift-stacking and other co-additive methods). With only two members of the interstellar small body population, this science use case relates to “unknown unknowns" since more classes of interlopers may exist. Interstellar objects on highly eccentric ($e \gg 1$) orbits are expected to be more difficult to detect than objects on trajectories that somewhat resemble those of long-period comets. ‘Oumuamua’s eccentricity was e $\sim$ 1.2, and Borisov’s eccentricity was e $\sim$ 3.5.

B.7.3.4. *Running on LSST Datasets (for the first 2 years)*—The wide-fast-deep strategy by Rubin/LSST will be ideal for detecting large numbers of small bodies. Especially because orbits of ISOs are hyperbolic, the initial stages of the LSST survey should rapidly yield a number of these objects. When it first surveys the sky, LSST’s deep limiting magnitude compared to other wide-field Southern Hemisphere observatories may result in the immediate detection of objects that are passing through this increased volume. New objects would be detected continuously as the small bodies enter the survey volume during the ten-year survey. Therefore, the greatest marginal increase in the scientific understanding of the ISO population will occur during the first few years of LSST operations. Population-level studies on the number of interstellar small bodies could be conducted through data releases and would not be subject to immediate time constraints. However, each individual the interstellar objects themselves will demand immediate follow-up for compositional and dynamical characterization.

This science case does not require additional science products from LSST, although it will need for selection functions and detection algorithms to be quantified for an extended parameter space compared to Solar System small bodies. In addition, an optimized post-processing pipeline will be important to efficiently evaluate population-level ramifications of LSST findings. Although the composition of ISOs should not affect detectability as much as these objects’ orbits, it will be important to understand this complication.

B.7.3.5. *Precursor data sets*—Similar algorithms could be validated using Pan-STARRS small body detection data, although that research might require its own equivalent of the in-development Solar System small body post-processing pipeline. Because of the dramatic effect which LSST will have on the catalog of ISOs, no precursor dataset can parallel the forthcoming data. Therefore, it is possible that efforts are best spent on developing well-understood simulated LSST datasets of ISOs.

B.7.3.6. *Analysis Workflow*—Orbital parameters of identified ISOs and small body selection functions corresponding to their orbital properties will be obtained from LSST data sources. Selection functions should decouple orbital data from the object's position on the sky and position on the the Rubin Observatory detector from the small body's inferred physical attributes. These data will be coupled with forward models that inject synthetic sources into survey images to determine the detectability of ISOs in real LSST frames. From this research, the interstellar object number density can be quantified for the Solar neighborhood.

Under the assumption that the region surrounding our Solar System is representative of the broader galactic interstellar object reservoir, one can extrapolate to the entire Milky Way. Simulations of small body ejecta from young exoplanetary systems can be run at any time before, during, or after the LSST survey and can be done on computing systems outside of the collaboration.

B.7.3.7. *Software Capabilities Needed*—Much of the software capability to complete this science case will build upon general software for small body selection functions. Moreover, this science use will directly benefit from the development of both tracklet-based and non-tracklet small body identification algorithms. In addition, readily available orbital data from LSST, potentially in conjunction with the Minor Planet Center, would be helpful for population-level research.

B.7.3.8. *References for Further Reading*—Do et al. (2018), Portegies Zwart et al. (2018), Engelhardt et al. (2017)

B.7.4. *Multiwavelength studies of Solar System moons and asteroids*

Contributors: Ilhuiyolitzin Villicana Pedraza

B.7.4.1. *Abstract*—Multiwavelengths studies have been important in characterizing new discoveries across astrophysics (e.g. Villicaña-Pedraza et al. 2017a). We expect the similar to be the case in the Solar System. We will analyze the dataset of ∼1 billion measurements from the Rubin Observatory's catalogs, combine them with light curves and spectra and multiwavelength information coming from instruments such as Atacama Large Millimeter Array (European Southern Observatory (ESO)) (ALMA), VLA, and JWST.

B.7.4.2. *Science Objectives*—The aim of our work is to support the analysis and observations of small bodies of the Solar System using the Rubin Observatory, and complement these with observations in the radio and the infrared. Using the Rubin Observatory we can cover a large portion of the bodies in the Solar System, using ALMA and VLA we can obtain important molecular spectroscopic information (Villicaña-Pedraza et al. 2017b) and using the JWST we obtain the IR data.

B.7.4.3. *Challenges (what makes it hard)*—Obtaining new observations from highly-oversubscribed telescopes is always difficult. We can also search information from archival data, catalogs, and the literature. Getting access to follow-up telescope time will be a challenge.

B.7.4.4. *Running on LSST Datasets (for the first 2 years)*—Data release catalogs LSST from data products will be analyzed. For the light curve we need 30 points for every object. We are planning to create catalogs for the LSST as products. We estimate that the light curves to be analyzed are 10. We can use the Asteroid Light curve Photometry Database[10] to supplement the light-curve information.

B.7.4.5. *Analysis Workflow*—First we need compare the public catalogs for asteroids to avoid duplication. We will use these catalogs to compare with our observations to match the existing data sets. The non-LSST information will be analyzed using tools such as DS9 for imaging and IRAF for spectroscopy. For ALMA and VLA data we will use Common Astronomy Software Applications (for ALMA) (CASA). These external data will need to be cross-matched to the LSST sample[11].

B.7.4.6. *Software Capabilities Needed*—We will use Rubin Data Management and LINCC software for the LSST. For ALMA and VLA we will use CASA. We will access other datasets or cross-match to other catalogs like https://alcdef.org/ for asteroids.

[10] http://alcdef.org/

[11] An example of cross-matching with present-day tools can be found at https://whitaker.physics.uconn.edu/wp-content/uploads/sites/2038/2017/02/PythonTutorial_Ashas.pdf

B.7.5. *Small Bodies in Rubin/LSST Data for Population-Level Studies*

Contributors: W. Garrett Levine (garrett.levine@yale.edu), Henry Hsieh (hhsieh@psi.edu)

B.7.5.1. *Abstract*—Conducting population-level studies of small bodies can illuminate the ancient Solar System's dynamical history. The orbital evolution of the giant planets sculpts the statistical distributions of small bodies, making these objects a viable tracer of the Solar System's assembly. We discuss a modeling pipeline through which users can impose physical parameters on synthetic sets of small bodies and simulate the resulting observations from Rubin/LSST. In addition, this pipeline can be used to simultaneously fit the physical parameters of small bodies to brightness data in LSST. Generally, the brightnesses of small Solar System objects depend on multiple physical parameters including heliocentric and geocentric distances, phase angles (i.e., the Sun-object-observer angle), rotational phase, viewing aspect angle (i.e., orientation of the object from the point of view of the observer), and filter. Phase functions expressing the brightness dependence of an object on phase angle and rotational lightcurves expressing the brightness dependence of an object on rotational phase are typically solved on their own, but our novel approach solves these parameters simultaneously. Because this pipeline is efficient, it keeps pace with the nightly alert stream and can be readily applied to simulate the small body yields and characteristics from LSST.

B.7.5.2. *Science Objectives*—Rubin/LSST will detect millions of Solar System small bodies with a diversity of orbital and physical properties. Determining the composition of these objects is necessary to advance models of planet formation and to constrain the dynamical history of the ancient Solar System. Broadly, we can divide this science case into two deliverables. First, it will be necessary to develop a module that computes best-fit properties (size, shape, composition, phase function, and rotation period) from input data from LSST (a series of times of observation, RA, dec, apparent motion, and calibrated brightnesses in multiple filters). For efficiency purposes, physical and orbital properties could be decoupled.

To conduct population-level studies with this tool, it will be necessary to finalize development of the Solar System small body post-processing pipeline to enable timely simulations of realistic Rubin/LSST datasets. Modeling small body populations via injection/recovery testing will be crucial. The aforementioned joint-fitting capabilities for color, rotation, and phase function will be important. Below, we provide a few examples of specific questions that may be addressed by this more general science case.

1. Connecting the elongation of monoliths (<100m objects) to the collisional physics which generated these bodies.
2. Estimating the population of 'Oumuamua-like (elongated) interstellar objects.
3. Comparing the intraclass and interclass characteristics of collisional families.
4. Identifying outliers in groups of asteroids with similar orbits.
5. Characterizing trends in shapes or rotation periods as a function of orbital elements and sizes

B.7.5.3. *Challenges (what makes it hard)*—Currently, a holistic model of asteroid brightness that accounts for the relevant physical parameters has yet to be integrated with the Rubin/LSST post-processing pipeline. Common routines that fit the parameters of small bodies often fit each attribute independently, which may lead to internal inconsistencies for the outputs. Joint-fitting routines can be algorithmically expensive, so computational efficiency in the Solar System post-processing module will be critical to the success of this science case.

Solar System objects may fall on different parts of the detector over different nights and possibly within the same night, especially for fast-moving and nearby small bodies. Therefore, well-constrained pixel-by-pixel detection efficiencies and selection functions will be important for this science case. Small bodies will move relative to background stars – accurate photometry for these objects will rely on high-quality difference image processing, or will otherwise need to be disentangled from any imperfectly subtracted background sources or image artifacts. Finally, the detectability of objects with extreme large-amplitude lightcurves (e.g., binary systems or objects with highly elongated shapes) that are close to the LSST detection limit may be highly dependent on the rotational phase at which they happen to be at the time of any given observation.

B.7.5.4. *Running on LSST Datasets (for the first 2 years)*—In order to identify scientifically valuable objects that require immediate follow-up, preliminary fits for physical parameters should be performed using photometry from the nightly alert stream once a certain amount of data judged to be sufficient for deriving those parameters with meaningful precision and reliability, where the threshold for data set size per object has yet to be determined. As more data are collected, these fits could be updated at regular intervals, where the ideal incremental increase in data set size needed to trigger re-fitting has similarly yet to be determined. Finally, given that some parameters may be expected to change over time due to changes in viewing aspect angle (which can cause an object's average projected cross-section on the sky to change, changing its apparent average size), resurfacing due to close encounters with other bodies or otherwise undetected collisions that can change an object's observed average color, or radiative forces that can accelerate or decelerate an object's rotation, the capability to perform physical parameter fits on Solar System objects using subsets of the full LSST data set would also be very desirable. Inactive (or marginally active) objects could be evaluated from small postage stamps, but active small bodies may require larger cutouts. More detailed fitting and population-level studies could be conducted through annual data releases.

B.7.5.5. *Precursor data sets*—Algorithms to predict asteroid brightness from a set of physical parameters could be tested on Pan-STARRS data along with selective follow-up of some small bodies to confirm the validity of results. When LSST is running, objects that have previously appeared in Pan-STARRS data could be cross-matched to assign longer baselines. Cross-calibration of the photometry would be necessary for executing this idea.

B.7.5.6. *Analysis Workflow*—Once a certain minimum amount of data (to be determined) has been acquired for a given Solar System object, those data should be processed using the algorithm(s) for simultaneous fitting of physical properties (primarily color, phase function, rotation period, and axis ratio). The most interesting and potentially scientifically valuable objects should be flagged for follow-up observations, and these targets should be disseminated to the Solar System Science Collaboration or to designated TOMs. In addition to storing the results of physical parameter fitting procedures in the LSST or broker databases, it will also be essential to record the precise data set used to derive those parameters to inform algorithms for triggering updated fitting analyses when the available data set for an object has increased by a sufficient amount that re-fitting is likely to produce meaningfully more precise results. Mechanisms should also be put into place to regularly perform fitting analyses only using subsets of the total LSST data set (e.g., all data in a given year or a given observational apparition, or every N consecutive detections, for example) to enable relatively unbiased searches for changes in physical parameters over time.

More computationally-expensive algorithms could be run on annual data releases. Any population-level analysis involving the post-processing pipeline could also be run on the data releases, since this work would not be time-critical.

B.7.5.7. *Software Capabilities Needed*—A model of small body observables (brightness, variability, color) must be linked to physical properties (size, shape, rotation). This code should be integrated into the post-processing module for Solar System objects. Since each small body is independent, this problem can be considered "embarrassingly parallel." Various algorithms have been developed for simultaneous fitting of some physical parameters (see References for Further Reading), but more evaluation is needed to determine which (if any) of these algorithms are best-suited for achieving our scientific goals and how much further development (if any) is needed to be able to operate them at LSST scales. It would also be highly desirable to have automated mechanisms to continuously monitor and evaluate the data available for every Solar System object in the LSST database to determine when both initial physical parameter fitting and updated fitting analyses should be triggered, where appropriate triggering criteria are likely to be somewhat more sophisticated than simply the number of available data points (e.g., in particular, also taking into account phase angle coverage).

B.7.5.8. *References for Further Reading*—Kaasalainen & Torppa (2001) "Optimization Methods for Asteroid Lightcurve Inversion. I. Shape Determination", Icarus, 153, 24
Lindberg et al. (2022) "Characterizing Sparse Asteroid Light Curves with Gaussian Processes", Astron. J., 163, 29
Lu & Jewitt (2019) "Dependence of Light Curves on Phase Angle and Asteroid Shape", Astron. J., 158, 220
Thirouin et al. (2016) "The Mission Accessible Near-Earth Objects Survey (MANOS): First Photometric Results", Astron. J., 152, 163

Waszczak et al. (2015) “Asteroid Light Curves from the Palomar Transient Factory Survey: Rotation Periods and Phase Functions from Sparse Photometry”, Astron. J., 150, 75

B.7.6. *Shift-and-Stack for faint object detection*

Contributors: Mario Juric, Steven Stetzler, David Trilling, Andy Connolly, Hayden Smotherman

B.7.6.1. *Abstract* —A foundational goal of the LSST is to map the Solar System small body populations that provide key windows into understanding of its formation and evolution. This is especially true of the populations of the Outer Solar System – objects at the orbit of Neptune and beyond (further than 30 AU). LSST, on its own, will detect individual KBOs to r~24.5. But advanced shift-and-stack algorithms (e.g., such as Whidden et al. 2019) would enable significantly deeper searches. This would allow for a census of the outer Solar System (OSS) to deeper magnitudes. For example, stacking just 10 epochs would reach ~25.5 magnitude, yielding 4-8x more TNO discoveries than the single-epoch baseline, enabling rapid identification and follow-up of unusual distant Solar System objects in $\gtrsim 5x$ greater volume of space (Jurić et al. 2019). Stacking O(36) exposure (~ half a year) would reach ~26.5, potentially yielding as many as 10^6 KBOs (extrapolating the early results of the DEEP survey; Trilling et al., in prep). These increases would enhance the science cases discussed in the Schwamb et al. (2018) whitepaper, including probing Neptune's past migration history as well as discovering hypothesized planet(s) beyond the orbit of Neptune (or at least placing significant constraints on their existence).

B.7.6.2. *Science Objectives* —

- Exploratory analysis of TNO populations across the entire sky to 25.5+ magnitude.
- Improve the characterization of OSS populations (size and color distributions at the small-size end).
- Increasing the number of tracers of dynamical populations in the OSS by 5-10x.
- Increased likelihood for discovery of distant or unusual objects
- Prior art: the work in progress on the DEEP survey (w. DECam), papers by Whidden et al. (2019) and Smotherman et al. (2021).

B.7.6.3. *Challenges (what makes it hard)* —

- Need access to (all) images (eventually) instead of catalog. Access patterns similar to building coadds, but requires all pixels in a ~3deg radius of a typical boresight. Significantly more expensive to produce the "shift and stack coadds," though the output dataset is small, ~O(TBs).
- Fast searches require access to accelerated computing hardware (GPU).
- Software is required to interface with a GPU both efficiently and with ease for a user.
- **Significant** algorithmic and implementation improvements needed:
  - Run time scales with the number of images searched. In the first year of LSST, we expect 80 images available to search. Based on a 24.5mag 5-sigma detection threshold on a single exposure, 7 images are needed for a 25.5mag search, 40 images for 26.5mag search, and 80 images allows for a search to 26.9mag. These numbers scale the run time by 10-100x.

  - Run time also scales with the time baseline searched, due to the search parameter space expanding. With present-day algorithms, the number of trajectories searched scales with the square of the time baseline. A rough calculation implies ~500 GPU-days to perform a search on 80 images over a 3 month baseline. This becomes ~63 days with 10 images, and ~7 days with 10 images over a 1 month baseline.
  - Current implementation needs to fit the required pixels into limited GPU memory. This imposes a constraint on the quantity (time baseline)x(sky area searched) since an increased time baseline increases the number of images and the sky area searched increases the number of pixels per image used. Assuming a stack of 80 images, cutting out a (1/16) deg^2 region of the sky requires ~36 GB of memory, reaching the memory limits of current GPUs.
  - Finally, the run time scales with the area of sky searched. The calculation above shows we can process (1/16) deg^2 of the sky with one GPU. Taking the size of the ecliptic to be 360deg × 10deg = 3600deg^2, this implies that – assuming just present day codes were used without algorithmic and implementation improvements – a full-sky search would take 57,600 GPU-years to complete.

- The need for repeated runs – analysis would be run ~quarterly-yearly, eventually over 10yr dataset.

B.7.6.4. *Running on LSST Datasets (for the first 2 years)* —

- Either raw/calibrated/difference LSST image dataset (both WFD and deep-drilling). The analysis can start as soon as a ~month of data is collected, but will likely be done on a quarterly or an annual basis (depending on the speed of the algorithm).
- The input dataset is the amount of single-epoch data LSST will collect over the stacking window (e.g., 3 months or 1 year).
- The output data size is small O(few TB).

B.7.6.5. *Precursor data sets* —

- These searches are already being performed with data from the DECam: High Cadence Transient Survey (HITS) survey, Deep Extragalactic Evolutionary Probe (DEEP) survey, and the deep drilling fields of the DECam Alliance for Transients (DECAT) survey (a direct precursor to LSST DDFs)

B.7.6.6. *Analysis Workflow* —

- For a given (set of) test orbit(s), the overlapping set of LSST visits will be determined. This will be ~30-90 nights worth of visits.
- For each LSST visit, a difference image exposure (data/mask/variance planes) will be obtained from the LSST archive and loaded into GPU memory. Images will be additionally masked to mask variable stars from the images.
- The GPU search will be performed, producing a set of candidate trajectories for moving objects Whidden et al. (2019).

- Postage stamps of the calibrated images will be queried from the LSST archive for each visit along a candidate trajectory. Forced photometry will be performed on the cutout image using LSST pipeline code.
- Postage stamps are validated using both human vetting and machine learning methods (a CNN trained with simulated postage stamps and previously discovered real objects).
- Validated detections are fit with an orbit to produce a discovery.

B.7.6.7. *Software Capabilities Needed —*

- Ability to query the LSST archive for calibrated exposures, difference images, and cutouts.
- New software infrastructure will be required to run this analysis, likely requiring many (10s-100s) machines with 1 or more GPU(s).
- Derived data products will include test orbits corresponding to likely stacked detections.

B.7.6.8. *References for Further Reading —*

- Shift-and-stack with GPUs: "Fast Algorithms for Slow Moving Asteroids: Constraints on the Distribution of Kuiper Belt Objects", Whidden et al. (2019).
- Extra depth from shift-and-stack with LSST data and KBO discovery estimates: "Enabling Deep All-Sky Searches of Outer Solar System Objects", Jurić et al. (2019).

B.8. *Cosmology*

B.8.1. *Weak lensing cosmology analysis / cosmic shear*

Contributors: Rachel Mandelbaum (rmandelb@andrew.cmu.edu), Arun Kannawadi (arunkannawadi@astro.princeton.edu), Andresa Campos (acampos@cmu.edu)

B.8.1.1. *Abstract* —We present the first cosmological analysis of the weak lensing shear-shear correlation functions, which trace the evolution of large-scale structure from intermediate to low redshift. With the first year of LSST data, we can already constrain the amplitude of matter fluctuations to greater precision than precursor surveys such as Kilo-Degree Survey (KiDS), DES, and HSC, reaching ~1% precision when considering statistical and systematic uncertainty. This improved result is due to the increased area of LSST, and the fact that the first year of data is comparable in depth to DES and deeper than KiDS. This work will rely on analysis of the catalogs produced by the LSST Science Pipelines down to detections with ~10σ significance. This result is sufficiently precise to permit a robust test of a Λ Cold Dark Matter; cosmological model (LCDM) by comparing with the amplitude of fluctuations from the CMB, which current analyses suggest may be higher than that from weak lensing (implying some uncontrolled systematic or a failure of the LCDM model). Together with the CMB data, we can go beyond the LCDM model and provide competitive constraints on the equation of state of dark energy.

B.8.1.2. *Science Objectives* —Milestones to this result include defining samples that pass basic null tests, validating that the shear estimates meet the needs for the science, and estimating the ensemble redshift distribution N(z) for the samples selected based on photo-z. For the last step, we do not necessarily need very precise photo-z (though that would improve the localization of the selected redshift bins and improve the constraint on w) but we do need a very precise understanding of the N(z) for the ensemble. At this stage, we anticipate that photo-z calibration and the impact of blending on shear and photo-z are likely to be the key limited systematics.

B.8.1.3. *Challenges (what makes it hard)* —Computational challenges related to image analysis, shear calibration, and cosmological likelihood analysis tend to compete with building sufficient models for systematics as the biggest challenge.

Broadly, there are two kinds of systematics: modeling systematics that are mostly astrophysical in nature and systematics in our data. Understanding astrophysical systematics that arise from our lack of understanding of baryonic physics and galaxy formation relies on information beyond the LSST data to mitigate their impact on cosmological parameter constraints. While some amount of self-calibration may be possible, we need to improve external (beyond LSST) measurements and simulations that give us direct constraints on baryonic feedback, intrinsic alignment etc.

Key systematics of both types are listed below:

- intrinsic alignments of galaxy shapes (IA): At LSST precision, modeling the intrinsic alignments of galaxy shapes with the large-scale density field is necessary. They are

one of the major drivers of the overall uncertainty in current cosmic-shear analyses. We will need tighter priors from external measurements and/or from hydrodynamical simulations. Self-calibration of IA parameters from the observations is also a possible approach.

- Blending: At LSST depths, a substantial fraction of the galaxies are blended. Nourbakhsh et al. (2021) estimates 12% of LSST galaxies would constitute undetected blends, i.e., two overlapping galaxies In general, they could have vastly different redshifts which necessitates a deblending algorithm that can apportion the flux to the different galaxies. This is currently done using the SCARLET algorithm (Melchior et al. 2018), which uses multiband images to model out the neighboring galaxies. Synthetic source injection tools that are developed jointly between DM and DESC can be used to estimate the sensitivity of the measurement to the amount of blending present. On longer time scales, pixel-level joint processing of Euclid+Rubin images to create Euclid+Rubin DDPs will provide an even better handle on deblending, especially undetected blends.
- PSF estimation: Zhang et al. (2022) estimates the bias in the cosmological parameters from errors in higher-order moments. They find that errors from the current version of the state-of-the-art algorithm Piff (now default in Rubin Science Pipelines) are comparable with PSFEx and that they introduce biases that are 0.2σ in the structure growth parameter S_8 in LSST Y1 analysis.
- Shear calibration: Even in the absence of any blending, source detection algorithms may preferentially detect sources that are sensitive to ellipticities/shape of the source. This introduces a detection or selection bias in the sample (Kannawadi et al. 2019). This could be corrected for by introducing a detection step in the shear measurement, which is essentially what metadetection achieves (Sheldon et al. 2020). The requirement on the shear multiplicative bias is at the level of 10^{-3}. While shear measurement algorithms such as metadetection reach this level in simulations with a number of realistic effects (E. Sheldon, internal communication), this could go beyond the requirements if various uncorrected systematic effects add up coherently. Can we get shear measurement codes to the stage where they do not need any image simulations for calibration (only validation)? It would be an ever-increasing burden to calibrate shear if calibration simulations are a requirement.
- Photo-z calibration: Blending impacts the ability to precisely estimate ensemble redshift distributions, which can ultimately cause a bias in the cosmological parameters, as shown in Nourbakhsh et al. (2021). Although we expect to have a different deblending methodology from the one in the above paper, it shows the importance of being able to perform deblending accurately.
- Sociological systematic: True blinding is difficult to achieve since the Rubin data products are available to the entire data rights community. DESC products can be blinded, but that would necessarily imply that Rubin data products cannot be used off-the-shelf for cosmological analysis by DESC.

These analyses will be done with a subset of the LSST data releases. The analyses will be mostly constrained by the major validation effort that is required before getting the results published, which has meant that most precursor surveys do not carry out cosmological shear measurements with each new data release.

B.8.1.4. *Existing Tools* —

- The following tools already exist today for this analysis
  - RAIL+qp for getting a p(z) for each object
  - Metadetection algorithm (descwl_coadd, mdet-shear-sims. . .)
  - Code from Pedersen et al. on self-calibrating IA
  - Synthetic Source Injection (DM/DESC)
  - Theoretical modelling (e.g., CCL, HMCode/Halofit), correlation functions (e.g., TreeCorr, NaMaster, TXPipe), Covariances, Inference (e.g., CosmoSIS)
- The following functionality is still missing from these tools:
  - Storing the p(z) for each object from different colors could be memory-intensive. We need either an uber-fast algorithm that can take in colors and reference catalogs and spit out p(z) on-demand, or find an effective compression technique to store them.
  - Current Synthetic Source Injection (SSI) software is not equipped to inject sources onto coadds, since the ability to inject a source and additional noise that is consistent with the source is not yet present. Developing this ability could save a lot of processing time.

B.8.1.5. *Running on LSST Datasets (for the first 2 years)* —We will use the following data products:

- Data release catalogs for the Wide Fast Deep (WFD) survey - we will use essentially all galaxies down to some relatively low significance (~10-sigma).
- Some calibration and systematics tests will use the DDFs.
- External spectroscopic catalogs to validate photometric redshifts.
- We will need some image simulations (beyond DESC DC2) that mock-up the cell-based coaddition to validate the shears, and cross-matching and cross-correlation against spectroscopic surveys to validate the photo-z and ensemble N(z) estimates.
- Euclid+Rubin DDPs for deblending (Schuhmann et al. 2019 predict mean shrinkage in ellipticity error bars by more than a factor of 2 when LSST is combined with Euclid)

B.8.1.6. *Precursor data sets* —

- DES, HSC, and KiDS all have public data releases that could be used for precursor analyses. In some cases, the catalogs are most accessible while reanalyzing the images would be a substantial challenge despite data being made public.

- Reference spectroscopic samples (deep ones for direct calibration and wide ones from e.g. DESI for cross-correlation) will be important for the N(z) validation.

B.8.1.7. *Analysis Workflow*—Assuming the existence of catalogs with galaxy shape or shear measurements and with photometric redshifts, the analysis workflow is as follows:

1. Apply selection criteria to the catalogs following some pre-defined process to mitigate selection biases (potentially using the metadetection technique).
2. Apply any shear calibration or other steps.
3. Divide the sample into tomographic (redshift) bins based on the photometric redshift estimates.
4. For each tomographic bin, apply a method for inferring the ensemble redshift distribution, N(z), based on the photometric redshifts and/or other information such as clustering with a reference spectroscopic sample.
5. Measure various null tests designed to detect the impact of systematic biases; depending on the results, may need to iterate on the above steps.
6. Measure shear correlation functions using pairs of tomographic bins i and j.
7. Produce a covariance matrix using some method potentially based on numerical integration of analytic expressions or based on mock catalogs.
8. Carry out a likelihood analysis for the cosmological parameters based on the measured data vectors and covariance matrices, including models for astrophysical and other systematics.

B.8.1.8. *Software Capabilities Needed*—Needed capabilities include the following:

- Survey property maps (e.g., skyproj with healsparse[12])
- Rubin calibrated exposures and coadds
- Estimators for ensemble redshift distributions

B.8.1.9. *References*—Precursor survey analyses:

- KiDS-1000: Asgari et al. (2021)
- DES-Y3: Amon et al. (2022)
- HSC-Y1: Hamana et al. (2020)

[12] https://github.com/LSSTDESC/healsparse

B.8.2. *Probabilistic Type Ia supernova cosmology analysis*

Contributors: Alex Malz (aimalz@nyu.edu), Rachel Mandelbaum (rmandelb@andrew.cmu.edu)

B.8.2.1. *Abstract* —Our scientific objective is to constrain the expansion history and cosmological parameters using photometric Type Ia supernovae from the first ~year of LSST data, in combination with precursor distance ladder measurements. In particular, we would like to go beyond the traditional analysis methods, by self-consistently incorporating probabilistic information as follows:

- Classification probabilities (e.g., with BEAMS)
- Host identification probabilities (e.g., with zBEAMS)
- Redshift probabilities – e.g., with Supernova Cosmology Inference with Probabilistic Photometric Redshifts (SCIPPR)
- Selection function
- Debiasing on the (probabilistic) absolute magnitude.

The probabilistic approach will permit the use of a larger sample in early LSST data compared to a non-probabilistic approach that requires secure follow-up in early LSST data.

B.8.2.2. *Science Objectives* —The key steps in the analysis:

- Alert broker provides classification probabilities and redshift posteriors for possible hosts for potential supernovae.
  - If using early (first year) data, it will be challenging to understand the selection functions which come with light curves correctly classified as Type Ia SNe.
  - We run our own classification using additional points in the light curve after the initial detection of the supernovae.
- We apply additional selection criteria to identify ones that can be used for science, and potentially get new host probabilities and redshift posteriors (but still need the originals because they go into the selection function).
- Light curve fitting with population-level debiasing function.
- Hierarchical inference of cosmological parameters factoring in probabilistic information for hosts, classification, and redshift (though none of the existing algorithms currently work at scale).

B.8.2.3. *Challenges (what makes it hard)* —The key technical limitation is that most of the steps in this analysis do not run at scale, while it's not clear how to do selection function quantification from the level of alerts, so algorithm development needs a lot of thought.

Early light curve fits will be particularly impacted by sparsity. Lack of redshifts from Rubin in the first year will be an issue - the brokers will get them from different sources and they will definitely differ from DESC photo-z.

B.8.2.4. *Running on LSST Datasets (for the first 2 years)* —The analysis will use:

- The alert stream.
- DESC value-added photo-z catalogs.

B.8.2.5. *Precursor data sets* —PLAsTiCC and ELAsTiCC data challenges can be used as a test-bed, but photo-z are not self-consistently generated with the light curves, especially for PLAsTiCC. ELAsTiCC results could potentially be used to determine the selection function (validate a method for doing so), but the sample may not be large enough. A preliminary study using Pan-STARRS Medium Deep Survey data has already been conducted (see Jones et al. 2018a).

B.8.2.6. *Analysis Workflow* —Running the analysis requires us to run these steps:

- Classify lightcurves using additional points after the initial detection.
- Quantify selection function - limited by basic algorithmic questions.
- Probabilistic lightcurve and absolute magnitude fitting.
- Use redshift uncertainties and classification probabilities for cosmology (BEAMS - does not work at scale), host identification probabilities (zBEAMS - simple assumptions, does not work at scale), redshift probabilities (scippr - open source, does not work at scale)
- Hierarchical inference at scale, combining all of the above components in a self-consistent manner.

B.8.2.7. *Software Capabilities Needed* —The needed software capabilities boil down to integrating and scaling up existing probabilistic approaches. There are also visualization needs for probabilistic data products connected to intuitive traditional plots, which could be implemented in the RSP.

B.8.2.8. *References for Further Reading* —

- Bayesian Estimation with multiple Species (BEAMS): Kunz et al. (2007)
- zBEAMS (BEAMS plus host identification uncertainty)- Roberts et al. (2017)
- Supernova Cosmology Inference with Probabilistic Photometric Redshifts (SCIPPR): open source code for incorporating probabilistic classification and redshift to cosmology inference, does not yet work at scale - Peters et al. (2018); https://github.com/aimalz/scippr
- Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC): The PLAsTiCC team et al. (2018); Hložek et al. (2020)

B.8.3. *Optimal spectroscopic follow-up algorithms for Type Ia supernova cosmology*

Contributors: Alex Malz (aimalz@nyu.edu), Emille E. O. Ishida (emille.ishida@clermont.in2p3.fr), Bruno Quint

B.8.3.1. *Abstract*—The Recommendation System for Spectroscopic Follow-up (RESSPECT) project is a COIN-DESC joint endeavour which aims to enable the construction of optimized training samples for LSST photometric SN cosmology. It takes into account a realistic description of the astronomical data environment and available follow-up resources. Early LSST data can be used to validate the ReSSpect pipeline, ensuring the system is able to scale to LSST requirements and helping define which categories of transients have the potential to boost performance of machine learning classifiers – if included in the training sample. The system requires daily updates on its data sets with every newly observed photometric point and the entire pipeline should be daily run in order to identify the optimal candidates for the subsequent night. As a consequence, significant computing resources are required to daily process all updated data which survives selection cuts. By construction, half of the objects selected for follow-up will be composed of SNIa and the other half will contain transients which are easily mistaken with SNIa by ML classifiers. The final queried spectroscopic sample, when used as a training sample, will enable optimum results from any supervised learning classifier - thus ensuring exploitation of purely photometric LSST light curves for SN cosmology.

B.8.3.2. *Science Objectives*—Cosmological analysis with LSST's photometric supernova sample will be essential because the spectroscopically confirmed (and even spectroscopically confirmable) sample will not be sufficiently large to improve upon the current status quo. In order to fully take advantadge of the information contained in LSST data, it is necessary to optimize the exploitation of purely photometric light curves.

Photometric supernova cosmology requires classification of lightcurves, primarily done by machine learning algorithms conditioned on a training set. Currently, the only possible available option is to train on simulations or on the available spectroscopic samples. Although encouraging results have been obtained by using simulations as training, this method inevitably biases the results towards the already known – and modelled – light curve features present in templates. Alternatively, currently available spectroscopic samples are highly biased towards SNIa-like events, which leads to sub-optimal performance from ML classifiers. From the algorithm point of view, an ideal training sample should enclose examples not only of the class of interest, but also on all possible objects which can be mistaken by it. Our final goal is to enable purely photometric supernova cosmology by using RESSPECT to construct informative training samples for ML classifiers. In parallel, the construction of this highly informative sample will also enable further development in the modelling of astrophysical transients - in particular SN events which lie in the boundary of the Ia class.

B.8.3.3. *Challenges (what makes it hard)*—

- Feature extraction: the pipeline requires that all incoming data are represented as a homogeneous rectangular matrix. It currently transforms all data into features using a parametric fit. This procedure needs to be repeated every time a new photometric point is observed. This process is slow and scales fast with the number of objects which needs to be processed at each night.
- Retraining: The active learning loop requires the model to be retrained at each night after the data has been updated. Moreover, in order to use the cosmology metric to disentangle equally valuable batches of potential candidates, the training needs to be repeated multiple times for each batch. This procedure needs to be done before the subsequent observation night. As a consequence, the classifier inside the AL loop needs to be cheap and straight forward to train.
- Cosmology fit: the cosmology metric calculation is quite slow and is based on a simplistic model because anything more sophisticated would be even slower. Nevertheless, it still requires a lot of computational time to enable the cost calculation for a series of multiple batches.
- Updated substream of alerts: after the retraining is done, the updated ML model needs to be used to filter the original data stream (provided via one of the community brokers). This will require interaction between the pipeline and the broker interface, since the entire process should be automatized.
- Communication with follow-up facilities: once the list of candidates is identified, we need to automatically distribute the list to the available spectroscopic follow-up facilities. And once the labels are obtained they need to be included in the training. Ideally these processes would be automatic.
- Timing constraint: results cannot be put to use right away. What we learn from doing this will build the training set we use for photometric SN cosmology in early LSST data and inform follow-up strategies after year 1.

B.8.3.4. *Running on LSST Datasets (for the first 2 years)*—The analysis will use:

- The filtered alert stream (removing artifacts and already known transients identified via cross-match with publicly available catalogs)
- Optionally utilize DESC value-added photo-z catalogs
- Initial training requires a minimum number of spectroscopically classified objects (5 Ias and 5 non-Ia would be enough)
- Validation: requires set of complete light curves which could potentially be used for cosmology (pass quality cuts, pending classification)
- Spectroscopic classifications for appointed candidates: candidates would be known at each night and classification will be acquired using external follow-up resources and publicly available catalogs

B.8.3.5. *Precursor data sets*—Our initial tests show that it is not advisable to use precursor data sets for training. Starting from scratch with LSST data allows the construction of a tailored data set and avoid bias in classification results.

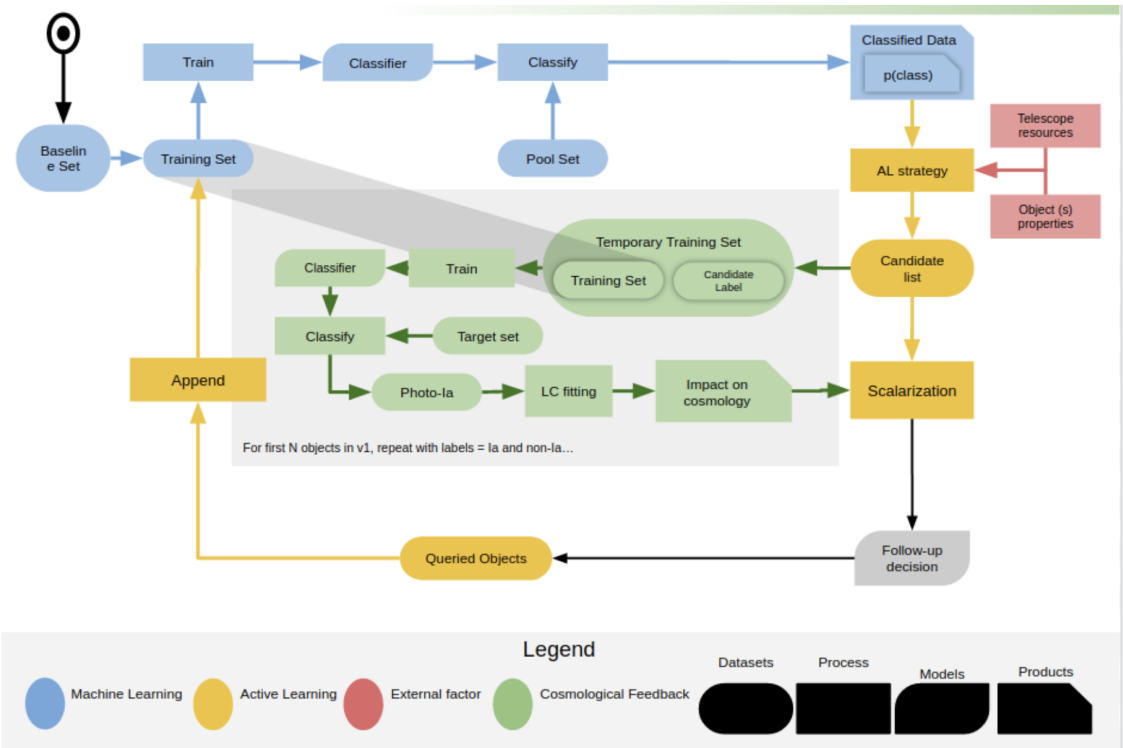

Figure 7. Workflow (figure courtesy of Bruno Quint)

B.8.3.6. *Analysis Workflow*—The key steps in the analysis (see also Figure 7):

- Construction full light curve (validation) and small spectroscopic confirmed (initial training) samples from alerts
- Apply feature extraction to the validation and initial training samples
- At each day (active learning loop):
  - Apply feature extraction to the pool and training sample including newly observed epochs
  - Train classifier
  - Classify test and pool sets (yields classification probabilities)
  - Identify candidates for spectroscopic follow-up
  - For each candidate, for each possible class:
    - Add the candidate with presumed label to the training sample
    - Train the classifier
    - Classify the test sample
    - Select photometrically classified SN Ia
    - Perform SALT2 fit (parametric SN lightcurve model) on photometric SN Ia sample
    - Add bias correction
    - Infer cosmological parameters (currently using a simplified Bayesian model due to computational expense)
    - Quantify cosmological impact for the new candidate in training set

- Combine cosmology metric of candidates into batches
- Evaluate observation cost of each batch
- Choose batch of candidates optimizing the overall metric
- Stream candidates to spectroscopic follow-up facilities
- Add new available labels (candidates from previous nights) to training set

B.8.3.7. *Software Capabilities Needed* —We have all the moving parts in terms of the software described above, and the pipeline works end-to-end on catalog data. We need an interface to the real alert stream as well as the spectroscopic follow-up facilities. We need software engineering to scale-up the computation (specifically the classifier retraining and cosmology metric evaluation) and the computing resources to run the pipeline nightly.

B.8.3.8. *References for Further Reading* —

- Basic active learning algorithm: Ishida et al. (2019)
- RESSPECT pipeline: https://github.com/COINtoolbox/RESSPECT/
- Cosmology metric validation code: https://github.com/emilleishida/resspect_metric
- Details on the active learning strategy and cost calculation: Kennamer et al. (2020)
- Active learning application to real alerts data: Leoni et al. (2021)

B.8.4. *Cross-correlation between LSST and CMB probes of gas physics*

Contributors: Giulio Fabbian (FabbianG@cardiff.ac.uk), Rachel Mandelbaum (rmandelb@andrew.cmu.edu)

B.8.4.1. *Abstract* —During its journey towards us, the CMB interacts with the large-scale structures of the universe (e.g. galaxy clusters, filaments, voids) as they form. This happens mainly through gravitational lensing of its photons and their inverse-Compton scattering with electrons having large thermal or bulk velocities (thermal Sunyaev-Zeldovich effect (tSZ) and kinetic Sunyaev-Zeldovich effect (kSZ)). Thus, the CMB is also a powerful tracer of the matter distribution in the universe. The matter distribution can also be probed surveying the distribution of galaxies at different redshifts and frequencies. The inhomogeneity of the matter distribution in the universe affects galaxy observations and CMB similarly, but, crucially, not exactly in the same way so that a significant amount of information can be extracted thus from their joint analysis and through their cross-correlations. Cross-correlations will help to break degeneracies between observables and to isolate different physical effects otherwise indistinguishable. The Simons Observatory (SO) will play a major role in this endeavor, providing a deep CMB polarization-sensitive survey covering about half the sky and the southern hemisphere starting in ~2023.

Multi-frequency observations of the CMB can be used to extract maps of the tSZ effect through its characteristic spectral distortion imprinted in the CMB, and to separate it from galactic foregrounds and extragalactic emission in the IR band. The kSZ signature is observed in maps of the CMB temperature anisotropies that cannot be easily isolated without the help of an external tracer given that its SED is the same as the one of CMB anisotropies. Alongside SO and on the same timescale, the next generation of galaxy surveys, the European Space Agency's Euclid mission and the Vera C. Rubin Observatory LSST, will map the mass distribution in the recent universe over more than half the sky with complementary experimental strategies. Given their full overlap, SO, LSST and Euclid will provide ideal data sets for joint studies and cross-correlation analyses.

B.8.4.2. *Science Objectives* —

- The cross-correlation between LSST galaxy density map and tSZ Compton y-map will give us insight on the bias-weighted pressure profile and energy content of electrons in halos.
- The cross-correlation between galaxy density of LSST and the square of the CMB temperature map (essentially a probe of the $\langle TT\delta \rangle$ bispectrum) is sensitive to both the electron density and velocity field (and hence a good probe of baryonic feedback processes).
- A good knowledge of the galaxy sample can be used to investigate the variation of both the physical processes above as a function of redshift and galaxy environment or mass.

- The analysis will use data products provided by CMB experiments and LSST catalogs with potentially augmented properties to allow for systematic tests of the sample selection and or investigation of variations as a function of environment properties.
- The accuracy of the characterization of the galaxy sample's ensemble redshift distribution (in a statistical sense) and extragalactic and galactic foregrounds affecting CMB data sets (e.g. modulation of sample selection function by extinction of galaxy dust) will be the major potential systematic factor. Additive systematics affecting only the CMB or the LSST data should not have a large impact in cross-correlation studies.
- Demonstrating analyses are currently being carried out with current generation of CMB data sets and publicly available surveys (that so far have been mainly optical spectroscopic surveys or IR photometric data, such as catalogs built from WISE data).

B.8.4.3. *Challenges (what makes it hard)* —

- Scientific challenge: The need for joint simulations to test the analysis methods.
- Technical challenge: Some cross-correlation software used for current surveys might not scale well to LSST data volume.
- Data quality:
  - Practical limit in resolution is imposed by CMB survey. Atacama Cosmology Telescope, for CMB observations (ACT)/SO give ~arcmin resolution y and T maps on potentially the full LSST footprint, SPT-3G will have lower noise but will cover a smaller sky area, while Planck has lower angular resolution and higher noise but potentially larger sky coverage. The details of the best data set to be used depends on the choice of the observing strategy and timing of the CMB experiments, i.e., which sky area will be covered (and delivered) first.
  - The need to have a reasonably well-understood redshift distribution for the galaxy sample used to define the density field could be a practical issue when defining the LSST galaxy sample.
- Timing issue: There is not a major urgency to this analysis, though carrying it out earlier rather than later might inform strategy and analysis plans for Cosmic Microwave Background Stage 4 (CMB-S4).

B.8.4.4. *Running on LSST Datasets (for the first 2 years)* —We'll use the photometric galaxy sample derived from object catalogs from 1 year of data (DR2). It will be important to have a set of well-characterized photo-z to define tomographic source samples for which we will infer ensemble redshift distributions. As long as the sample is reasonably homogeneous and inhomogeneities are well understood, the analysis can proceed.

We'll want a value-added galaxy catalog with additional characterization of the galaxy SEDs, local density, selection function, etc.

We'll also need a CMB dataset with sufficient resolution to carry out this analysis, such as ACT, South Pole Telescope (SPT), SO, etc.

B.8.4.5. *Precursor data sets* —Photometric catalogs from KiDS, DES, and/or HSC can be used to develop this use case though they cover a smaller area than LSST, so DES might be the only viable option. Overlap with a CMB dataset is a key practical consideration to define the details of the analysis. Joint simulations including realistic evolution of the dark matter and baryon distribution on both the galaxy sample as well as on the CMB probes will likely be important for carrying out this science with LSST and SO, as these probes are expected to be measured with high precision and statistical significance.

B.8.4.6. *Analysis Workflow* —

- Select galaxy catalog and split it in different mass or tomographic bins according to LSST redshift estimates.
- Make galaxy overdensity maps and store them.
- Compute and store cross-correlation statistics (usually angular power spectra) between galaxy overdensity maps and external CMB temperature and/or tSZ maps.
- Compute or estimate covariance from jackknife resampling or simulation or analytical approximations.
- Perform robustness or consistency/null tests using data splits constructed with subsets of the galaxy catalog or CMB products built from subset of the data depending on which data splits are available from the CMB side.
- Inference of relevant parameters.

B.8.4.7. *Software Capabilities Needed* —

- Query the LSST archive and make galaxy catalogs with various cuts and desired properties, including division into tomographic samples (distinct redshift bins).
- Map-making software, to construct maps of projected quantities such as galaxy overdensity at a variety of resolutions
- Ensemble redshift distribution inference for photometric samples.
- Cross-power spectrum calculation that takes maps as inputs.
- Theory predictions of 2-point statistics (through Core Cosmology Library, https://github.com/LSSTDESC/CCL (CCL) or similar software)
- Monte Carlo Markov Chain (MCMC) sampler or likelihood-free inference code [which would require fast simulations / new software development]

In some cases new methods or optimized software are needed but they are common needs for other cosmology science cases (e.g., 3x2pt analysis). Joint simulations would be more specific to this LSST+CMB science case.
Maps and power spectra points and covariances will be stored.

B.8.4.8. *References for Further Reading* —Example of prototype joint tSZ and kSZ analysis: (Schaan et al. 2021) (ACT and Planck CMB + Baryon Oscillation Spectroscopic Survey (BOSS) CMASS).
kSZ from Planck + DESI (with spectroscopic data): Chen et al. (2022).

kSZ from ACT + KiDS: Schneider et al. (2021).
tSZ from ACT and Planck + KiDS: Tröster et al. (2022).
SPT and Planck y-maps (Koukoufilippas et al. 2020).
Projected field kSZ (without spectroscopy): estimator and its application on data (Hill et al. 2016; Kusiak et al. 2021).
tSZ - galaxy density cross-correlation (Koukoufilippas et al. 2020).

B.8.5. *Self-consistent cosmological parameter constraints from galaxy clustering and galaxy-galaxy lensing using the DESI Y1 LRG sample*

Contributors: Kate Storey-Fisher (k.sf@nyu.edu), Francois Lanusse, Sam Schmidt, Arun Kannawadi

B.8.5.1. *Abstract*—We present the results of a galaxy clustering and galaxy-galaxy lensing analysis using the Rubin LSST photometric sample and the DESI spectroscopic sample. DESI is the current largest spectroscopic redshift survey, and is still ongoing. The two surveys have an overlap region of at least ~4000 deg^2, and could increase to as much as 6000 deg^2 depending on footprint decisions (Lochner et al. 2022). We use the DESI LRG sample, which contains ~8 million galaxies from $0.3 < z < 1$, as the lenses. We model galaxy bias, baryonic feedback, and intrinsic alignments, and the high precision of DESI allows us to break degeneracies to better handle these systematics. We constrain Ω_m to $X \pm Y$ and S_8 to $XX \pm YY$. This lensing analysis from combining LSST and DESI is an important cross-check of the LSST cosmic shear results; we find that [they give consistent constraints].

B.8.5.2. *Science Objectives*—To perform this analysis, will require Rubin's photo-z catalog and shape catalog for the LSST sample. We will also need the DESI spectroscopic galaxy catalog, of which there are already releases. The catalogs required for this joint analysis are also necessary individually for each survey's analysis, so this analysis would leverage the existing work already performed to improve constraints at little additional cost.

Our analysis will be limited by the size of the survey overlap region; the choice of extending the LSST footprint north further into the DESI footprint would allow for significant increased analysis power. To understand these limitations and make forecasts for the improvement possible with this joint analysis, we could perform a precursor analysis with other data sets of photometric sources and spectroscopic lenses such as KiDS and GAMA, or HSC and SDSS-III/IV.

B.8.5.3. *Challenges (what makes it hard)*—This joint analysis will require address a number of challenges, including:

- Understanding of clustering sample selection functions, but this should be handled by DESI. Would result in catalogs of random points illustrating the coverage as a DESI-provided data product.
- We will rerun the analysis for each Rubin data release; the challenge is to make the analysis easily re-runnable.
- The photo-z quality for Y1-Y2 LSST data will be lower than Y10, and the early LSST overlap with DESI may be less uniform in image quality (and the high airmass may be prioritized by the scheduler). Essentially, the depth maps may be more complex and have more structure early in the survey at the edge of the footprint.

- Most/all LSST source photo-z tomographic bins assume calibration with a DESI-like (probably DESI) sample; may need to worry about covariance when also using DESI for cosmology measurements.

B.8.5.4. *Running on LSST Datasets (for the first 2 years)*—We will require the following data sets and pay attention to relevant considerations as follows:

- Shape catalog: Metacal should be available from the DM pipeline, Metadetect may be available (otherwise would be a value added catalog from DESC).
- As the LSST/DESI overlap area is at higher airmass than the average LSST airmass, we will have to recalibrate the images in this region. As this region is ~1/5 of the entire LSST survey area, this will result in ~$\sqrt{5}$~2.2 times larger error bars. (Note that we will have to put in additional effort for the recalibration, which requires rechecking and potentially correcting for the multiplicative and additive biases of the weak lensing shears, but we will not have to fully re-do the shape measurements.). As g-g lensing is more forgivable in terms on calibration parameters, this may be tolerable. So a subset of the image simulations may be sufficient.
- Estimate of number of galaxies that can serve as source galaxies for DESI LRGs in early LSST data: 70 million. We obtain this from considering that in early LSST data, an estimated source galaxy number density is 10 arcmin^{-2}, of which approximately half are above the LRG redshift (The LSST Dark Energy Science Collaboration et al. 2018). Given the expected 4000 deg^2 overlap region, this gives approximately 70 million sources.
- Photoz for the shape catalog: Tomographic sample selection based on Rubin-provided photo-z, calibrated by cluster-z with DESI.

B.8.5.5. *Precursor data sets*—

- For the spectroscopic galaxy sample, we will be able to use early DESI data as a precursor. We could also use GAMA, a spectroscopic survey, or SDSS-III/IV.
- For the galaxy shape data set, we could use the Ultraviolet Near- Infrared Optical Northern Survey (UNIONS) survey, HSC, or KiDS.

B.8.5.6. *Analysis Workflow*—We lay out the following workflow as a possible route to completing this analysis:

- Perform an optimized tomographic binning of shape sample, likely using Rubin-provided photo-z estimates.
- Use DESI Emission-Line Galaxies (ELG)s & quasars for photo-z calibration.
- Obtain the DESI LRG catalog for the lens sample.
- Using the window functions of both surveys, create samples in the overlap region of the two surveys.
- Compute correlation functions (galaxy-galaxy lensing, as well as the auto-correlation) on these samples.

- Implement null tests. This might include, for example, checking that high-redshift lenses with low-redshift sources produce no signal.
- Estimate a covariance matrix using theory predictions or empirical estimators (jackknife etc.), accounting for correlations between the different quantities being measured.
- Carry out the cosmological parameter likelihood analysis given the data vectors, covariances, and models for systematic errors.

B.8.5.7. *Software Capabilities Needed —*

- We will require software to compute the correlation functions, such as TreeCorr[13] or a similar correlation function package. While we are working with large data quantities, we expect current techniques and computing power will be sufficient.
- We will need to store these correlation functions as intermediate data products, but this will not require a large amount of space.

B.8.5.8. *References for Further Reading —*

- Pandey et al. (2021): Dark Energy Survey Year 3 Results: Constraints on cosmological parameters and galaxy bias models from galaxy clustering and galaxy-galaxy lensing using the redMaGiC sample
- Zhou et al. (2020): Preliminary target selection of DESI LRG sample
- Bolton et al. (2018): Maximizing the Joint Science Return of LSST and DESI

[13] https://github.com/rmjarvis/TreeCorr

B.8.6. *Weak lensing cosmology analysis / 3x2pt*

Contributors: Andresa Campos (acampos@cmu.edu), Francois Lanusse (francois.lanussse@cea.fr)

B.8.6.1. *Abstract*—We present the first cosmological analysis of the weak lensing shear-shear correlation functions in combination with galaxy clustering and galaxy-galaxy lensing, which trace the evolution of large-scale structure from intermediate to low redshift. With the first year of LSST data, we can already constrain the amplitude of matter fluctuations to greater precision than precursor surveys such as KiDS, DES, and HSC, reaching ~1-% precision when considering statistical and systematic uncertainty. This improved result is due to the increased area of LSST, and the fact that the first year of data is comparable in depth to DES and deeper than KiDS. This work will rely on analysis of the catalogs produced by the LSST Science Pipelines down to detections with ~10-sigma significance. This result is sufficiently precise to permit a robust test of LCDM by comparing with the amplitude from the CMB, which current analyses suggest may be higher than that from weak lensing (implying some uncontrolled systematic or a failure of the LCDM model). Together with CMB data, we can go beyond the LCDM model and provide competitive constraints on the equation of state of dark energy.

B.8.6.2. *Science Objectives*—Milestones to this result include defining samples that pass basic null tests, validating that the shear estimates meet the needs for the science, and estimating N(z) for the samples selected based on photo-z. For the last step, we do not necessarily need very precise photo-z (though that would improve the localization of the selected bins and improve the constraint on w) but we do need a very precise understanding of the N(z) for the ensemble. At this stage, we anticipate that photo-z calibration and the impact of blending on shear and photo-z are likely to be the key limited systematics.

B.8.6.3. *Challenges (what makes it hard)*—Systematics! This science case directly inherits from the cosmic-shear systematics, we refer to Section B.8.1 for details, and here we document the additional galaxy-clustering challenges.

- Galaxy bias modeling: accurate modeling for galaxy bias used for galaxy-galaxy and galaxy clustering modeling will be needed. Ongoing DESC bias challenge is investigating various nonlinear bias models to reach Y1 and Y10 requirements.
- Baryonic Effects: feedback from stars and supermassive black holes (AGN) redistributes gas. The effects of galaxy formation are degenerate with cosmology/ neutrino effects, therefore ignoring baryons at small scales leads to biases (see Schneider et al. (2020)). It is necessary to model the baryonic effect and marginalize of it.
- Deprojection of systematics maps: to clean potential contamination from varying survey and observing conditions variability, a step of mode deprojection from a set of systematics maps can be applied (see e.g. Alonso et al. (2019)).
- Blinding: a blinding scheme that encompasses multiple probes is a challenge on its on, as discussed in Muir et al. (2020). Added to that, there is the additional question on

how effective the blinding of DESC data products will be, given that Rubin products will be available.

B.8.6.4. *Existing Tools* —

- What tools exist today to undertake these analyses
  - RAIL+qp for getting p(z) for each object
  - Metadetection algorithm (descwl_coadd, mdet-shear-sims. . .)
  - Code from Pedersen et al. on self-calibrating IA
  - Synthetic Source Injection (DM/DESC)
  - Systematics maps building (e.g. Supreme)
  - Theoretical modelling (e.g., CCL, HMCode/Halofit), correlation functions (e.g., TreeCorr, NaMaster, TXPipe), Covariances (TJPCosmo), Inference (e.g., CosmoSIS)
- What functionality is missing from these tools and frameworks that would need to be developed, or is there some issue with their application to the dataset at LSST scale?
  - Toolset is fairly complete for Y1
- If new tools were built what components from existing frameworks would be critical to keep? (e.g. what parts of existing tools work well)
  - Save All Correlations and Covariances (SACC) file format for interchangeable data files.

B.8.6.5. *Running on LSST Datasets (for the first 2 years)* —We will use the following data products:

- Data release catalogs for the WFD survey - we'll use essentially all galaxies down to some relatively low significance (~10-sigma).
- Some calibration and systematics tests will use the DDFs.
- External spectroscopic catalogs to validate/calibrate photometric redshifts.
- We will need some image simulations (non-DC2) that mock-up the cell-based coaddition to validate the shears, and cross-matching and cross-correlation against spectroscopic surveys to validate the photo-z and ensemble N(z) estimates.
- Euclid+Rubin DDPs for deblending (Schuhmann et al., 2019 predict mean shrinkage in ellipticity errorbars by more than a factor of 2 with LSST is combined with Euclid)

B.8.6.6. *Precursor data sets* —

- DES, HSC, and KiDS all have public data releases that could be used for precursor analyses. In some cases, the catalogs are most accessible while reanalyzing the images would be a substantial challenge despite data being made public.
- Reference spectroscopic samples (deep ones for direct calibration and wide ones from e.g. DESI for cross-correlation) will be important for the N(z) validation.

B.8.6.7. *Analysis Workflow*—This science case directly inherits from the cosmic-shear, galaxy clustering and galaxy-galaxy lensing workflow. We refer to Section B.8.1 and Section B.8.5 for details on each probe. Here we provide the general steps and document the additional items related to the combined 3x2pt analyses:

- Apply selection criteria to the source and lens catalogs.
- Apply the optimal redshift estimation method to each catalog, perform tomography and estimate the N(z) distribution.
- Measure each of the three two-point correlations: cosmic shear, galaxy clustering and galaxy-galaxy lensing.
- Compute the joint covariance matrix of the three probes.
- Model and mitigate systematics.
- Perform likelihood analyses and cosmological parameter estimation.
- Test for internal consistency of probes before unblinding.

B.8.6.8. *Software Capabilities Needed*—

- Survey property maps (e.g., skyproj w/ healsparse, Supreme)
- Calexps and coadds
- N(z)s estimator
- Sampler

B.8.6.9. *References*—Precursor survey analyses:

- DES-Y3: Abbott et al. (2022)
- KiDS-1000: Heymans et al. (2021)

C. TECHNICAL AREAS IN DETAIL

C.1. *Introduction*

After the science breakouts we identified six main cross-cutting technical areas, and developed use cases within them, with a goal of understanding the needs to support a broad range of Rubin science (Table 1). Each of these areas is given a subsection here.

In this section:

C.2. *Cross Matching*

Contributors: Viviana Acquaviva, Igor Andreoni, Leanne Guy, Saavik Ford, Nico Garavito-Camargo, Mario Juric (editor), Ilhuiyolitzin Villicana-Pedraza, Jeremy Kubica, Samuel Wyatt, Weixiang Yu, Alex Riley

C.2.1. *Abstract*

A significant fraction of science use cases presented at the Workshop require the ability to (generally positionally) cross-correlate the detections in the LSST catalog with one or more other catalogs – an operation commonly known as "cross-matching". This capability would enable enrichment of LSST data with information taken in other wavelengths, at other times, different resolutions, or of generally different characteristics. This capability is needed in two regimes: a) real-time – low-latency matching of O(10k) sources to O(10) catalogs each holding O(1Bn) objects (to support adding information to alert streams from other catalogs), and b) offline processing – the ability to match O(10Bn) x O(1Bn) object catalogs, followed by joining data from both catalogs (e.g., full time series of observations, multi-wavelength studies) for analysis at scale. This capability should be easy to use for the end-user. For example, it may be provided at the community broker level for real-time cases, or accessible as simple Python calls or SQL-like statements callable from Python notebooks for the offline-level.

We note that cross-matching is just a first step in the process, nearly always followed by bringing in additional data from catalogs being cross-matched. It may be better to think of it as *(distributed) joining of (large) tables on a spatial (user-defined) index, followed by (potentially heavy) computation on the result of the join*. The cross-match capability may therefore need to be a part of a larger scalable analytics system.

C.2.2. *Science Cases Needing this Tool*

Solar system:

- Multifrequency study of the Solar System moons and hazardous asteroids (Section B.7.4)

Local universe static

- The properties of the faint end of the Main Sequence: the stellar/sub-stellar boundary. (Section B.5.3)
- Mapping the Accreted and Intrinsic Stellar Populations in the Milky Way (Section B.5.1)
- Local Group Dwarf Galaxies, bound and unbound (Section B.5.2)
- The local IMF as inferred from nearby star forming regions and clusters (Section B.5.4)

Local universe variable & transient:

- All science-cases involving enriching discovered transients or variability with non-LSST data.

Extragalactic static:

- Estimation of galaxy physical parameters with SED fitting (Section B.2.4)

Extragalactic variable:

- Find All the AGN ASAP (Section B.4.3)
- Connection between short term variability of AGN and their long term behavior (Section B.4.4)
- Augmenting AGN variability (Section B.4.1)

Extragalactic Transient:

- Real time transient host association
- Immediate classification of astrophysical transients (Section B.3.1)
- Non-localized alert alert crossmatching

C.2.3. *Requirements for the software*

We identify two rough clusters of use-cases, with differing scalability and performance requirements:

1. **Real-time cross-matching:** The ability to return multiple (say, N<=10) neighbors within ~5", with latency at alert scale (seconds), and the numbers of objects being cross-matched similar to the numbers of alerts. The system well need to cross-match to multiple (arbitrary) catalogs with a choice of distance metrics (e.g., on-sky, real-space). Brokers may enrich alerts w. this information, therefore needing close collaboration with broker teams wishing to operate this functionality.

2. **Large-scale static-sky cross-matching:** The ability to do full catalog cross matches for statistical analysis. Matching to a filtered subset will be an important feature – for example, a query such as "for these 10M objects, find 3 nearest $g - r < 0.5$ neighbors". This system would need to:
    - Scale to O(10Bn) x O(1Bn) catalog cross-matches and repeat on ~monthly scales.
    - Allow matching as close to interactive as possible (e.g., with smaller catalogs) to enable exploratory science.
    - Have the ability to do faster cross-matching on smaller subsamples of these objects, for testing purposes.

Aspects that are common to both use cases include:

- Cross-matching is just a first step in the process, nearly always followed by bringing in additional data from catalogs being cross-matched. It may be better to think of this as (distributed) joining of (large) tables on a spatial (user-defined) index – a joint table of 100M objects is not very useful if the data on those objects (e.g., time series) can't be easily pulled/fused/processed together.
- Desirability of probabilistic cross-matching, which takes into account finite extent or a large uncertainty ellipse for some objects. This may be emulated by the user if a basic cross-matching capability where N nearest neighbors are returned is provided For example, the user can take N nearest cross matches, compute some probability function that each of them is the "true" match, and pick the most probable one (or keep the full distribution).
- N-way cross-matching (cross matching to multiple catalogs)
- Reporting non-detections is important. E.g., for Solar System, transients/variability, etc. Related to (reliable) selection functions.

Challenges:

- The low latency and partial scalability for the real-time (alerts) use case.
- The scalability for the online case, where large (likely distributed) catalogs are being joined. For example cross-matching the LSST (located at SLAC) and the DES catalog (located at NCSA), joining the time-series, and passing them on to a Python function to compute some metric of interest on the joined light curve.
- The join predicate which may be more complex than simple spatial join (e.g., "find N closest matches in a user-defined subset of a catalog").

C.2.4. *Running on LSST and other Datasets*

As discussed above, this component is expected to be needed both for real-time and batch processing use cases.

C.2.5. *Existing Tools*

- Codes such as the Apache Spark based AXS, for scalable distributed joining/cross-matching of extremely large datasets.

- Codes such as catsHTM and similar for fast in-memory crossmatching
- Online cross-match services like CDS XMatch and many others
- HEALPix Alchemy for non-localized events such as GW/Neutrino Alerts. Has the capability to crossmatch catalogs, observational footprints, and all-sky images within a healpix map.
- Cross-match services being developed as parts of community brokers

Needed enhancements:

- Scalability – many of these tools assume O(<1k-1M) object operation
- Need arbitrary catalog cross-matching, w/o pre-computed join tables and with user-defined predicates
- N-way cross-match – the ability to join many catalogs
- Distributed joins – many of these tools focus on just computing cross-match, but science use cases need to bring together and work on the data as the next steps (e.g., light curves, spectra, sometimes even images).
- User-friendly distributed operation – the user should be able to use these tools with similar ease to using a RDBMS (Relational DataBase Management System) today. For example, the user should be able to write a declarative statement about the result they're trying to achienve, and have the execution – query optimization, scaling, potential work distribution, fault tolerance – be handled transparently.

C.2.6. *Computational Workflow*

Computational workflows are somewhat science dependent, and many have been discussed in specific science use cases. Here we just give two representative examples:

Example from the stellar/sub-stellar boundary science case:

- Open a notebook and query some sources by making cuts (in color) on the object catalog data to identify candidate low-mass stars.
- Identify candidates with early proper motion and parallax measurements.
- Get sources data (light curves) and/or take source measurements for each selected object.
- Run routines to identify variability.
- Cross-match with calibration subsets (Gaia, spectroscopic surveys, etc.)

An example from the SED fitting science case:

- Open a notebook and query some sources by making cuts for example in redshift or luminosity.
- Run some object classification algorithm to identify which sources are galaxies.
- Cross match those galaxies with sources from other multi-wavelength catalogs to obtain a "wider" SED.
- Do science with the cross matched data (validation of LSST-results, SED fitting on multi-wavelength SED, comparison with simulations, and others).

- AXS: Astronomy eXtensions for Spark – Apache Spark based distributed cross-matching system (Zečević et al. 2019)
- HEALPix Alchemy (Singer et al. 2022)

C.3. *Selection Functions*

Contributors: Yusra AlSayyad, Katelyn Breivik, Giulio Fabbian, Matt Holman, Adrian Price-Whelan, Kate Storey-Fisher

C.3.1. *Abstract*

Selection functions are core components of any modeling procedure that aims to quantify the population statistics or density distribution of sources or objects. A selection function for a given modeling method may contain things like the detection efficiency of sources with a given brightness or shape, the classification accuracy of sources, the cadence of observations, the Milky Way and intergalactic dust distribution, or the crowdedness (in source counts) of a field. While the LSST Data Management (DM) group plans to provide the core data products, the detection efficiency of a theoretical point source per position and epoch and each coadd, each specific science case, classification, or detection algorithm will need a specialized selection function which depends on the science question or model being studied. What we are therefore currently missing in the community are worked examples of how to construct and use selection functions of varied complexity for different use cases.

C.3.2. *Science Cases Needing this Tool*

Any science cases that want to learn population statistics or source density distributions.

C.3.3. *Requirements for the software*

We recommend producing 3–5 science demonstrations that utilize the DM data products and metadata to construct selection functions that are used in illustrative science examples. As possible example use cases, we recommend:

- A selection function combining the depth of coadd images over the sky with a dustmap, to evaluate a model for the spherically-averaged distribution of stars in the Milky Way's stellar halo.
- Tools and worked examples on how to convert from the pixelized/rasterized representation of detection maps that DM will provide to vector/polygon representations for users who need them.
- Worked examples for generating simulations either at the catalog-level or realized images for injection into the Rubin images via the Pipelines' synthetic source injection framework.
- A selection function that incorporates cadence or time of observation information as a function of sky position to assess the detectability of RR Lyrae stars to measure the period and magnitude distributions of RR Lyrae stars. Similar for variability-selected AGN populations.
- A selection function for detecting asteroids of a given shape in single images to compare the numbers of asteroids with different shape characteristics.

C.3.4. *Running on LSST and other Datasets*

This requires LSST (coadd and single-epoch) image masks, property maps and both coadd and time-series catalog information. Specific tutorials will pull in case-specific info such as dustmaps and variability models.

C.3.5. *Existing Tools*

The core data products are planned to be produced by the DM team, but what is missing are demonstrations of how to construct selection functions of different types/forms.

C.3.6. *Computational Workflow*

n/a

C.3.7. *References for Further Reading*

Boubert & Everall (2022), Rix et al. (2021)

C.4. *Time Series*

There were multiple technical cases for time series.

C.4.1. *Parametric Fitting*

Contributors: Catarina S. Alves (mailto:catarina.alves.18@ucl.ac.uk), Matthew Graham, Andrew Bradshaw, Andrew Connolly, Garrett Levine, David Trilling, Fabio Ragosta, Tomas Ahumada, Jing Lu, Alex Gagilano, Neven Caplar, Illija Medan

C.4.1.1. *Abstract*—Many analyses involve fitting a prespecified model to data where the model parameters have semantic content, for example, they represent physical quantities. Given the numerous and diverse objects that Rubin LSST will observe with unprecedented precision and time coverage, including solar system objects, stars of all stripes, transients, and variable galaxy images, the models which are fit to the observations must similarly be flexible while also including physical information about each object. Commonly used tools in parametric time series analysis should be automatically computed (or trained) on all objects on a regular basis, along with specific subsets being analyzed with targeted tools as needed; all provided through a unified interface with a common data structure. Additionally, model selection criteria such as AIC or Bayes factors shall also be pre-computed to enable comparison between models, along with uncertainties and ranges of validity for model parameters. Providing the data alongside informative statistics in a networked and unified interface should help maximize the potential of LSST.

C.4.1.2. *Science Cases Needing this Tool*—Parametric fitting to time series is common to many areas of astronomy, including:

- cosmology (in particular, SN Ia cosmology to produce distance estimates)
- AGN to model stochastic variability
- stars to model particular periodic structure in light curves, e.g., RR Lyrae
- exoplanets to model background stellar activity

C.4.1.3. *Requirements for the software*—The primary requirements for the software are that it should:

- scale efficiently to arbitrary data sizes
- be flexible across different model structures and fitting constraints, e.g., optimizer choice, loss functions
- be able to provide uncertainties (posterior distributions) on both model parameters and predicted values
- run as fast as possible

In most cases, the input will be time series, either individual or batches, and these will be irregularly and sparsely sampled, gappy, and heteroskedastic. The time series may also be univariate (single filter) or multivariate (ugrizy). It may be that this is best supported through a common time series data model.

A lot of parameter fitting codes exist for specific science cases (see below) so rather than implementing new versions of these, the software requirements is to run these in an appropriate environment. Modification of existing codes might be needed to implement the uncertainty requirement and to improve the speed and scalability of the code.

C.4.1.4. *Running on LSST and other Datasets* —This technical case will use time-series, whose needed duration depends on the specific use case. The datasets could come both from calibrated images and data release catalogs. The alert stream data may present challenegs for parametric fits because the alerts use different templates across the years, which could result in artificial discontinuities.

SN Ia cosmology is a science case associated with parametric fits. There are numerous precursor datasets from previous surveys. More recently, we have data from the Zwicky Transient Facility (ZTF) and LSST-like simulated datasets such as the Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC; The PLAsTiCC team et al. 2018; Kessler et al. 2019) and its update, ELAsTiCC. The analysis of these precursor datasets will lead to publications, in particular when applied to real data. A comparison of template fitting algorithms on the same datasets, including the advantages and limitations of each methodology, would also lead to a new publication.

C.4.1.5. *Existing Tools* —

- Tools (SN Ia light curve fitting + SED templates + AGN): SALT2 (Guy et al. 2007; Taylor et al. 2021), SALT3 (Kenworthy et al. 2021), SUGAR (Léget et al. 2020), ParSNIP (Boone 2021), SNooPy (Burns et al. 2010), BayeSED (Han & Han 2012, 2014, 2018); EzTao (Yu & Richards 2022)
- There is no easy way to compare between the above tools. Moreover, the SALT light curve fitters are too slow for LSST use.
- It is crucial to work with the community to understand how to implement the unified interface with a common data structure to maintain the ease to use of the existing tools.

C.4.1.6. *Computational Workflow* —Pick your optimizer, pick your loss function. Compress data from photometry as a function of time into phase and amplitude.

C.4.1.7. *References for Further Reading* —Links to tools:

- *SALT2* (Guy et al. 2007; Taylor et al. 2021)
- *SALT3* https://saltshaker.readthedocs.io/ (Kenworthy et al. 2021)
- *SUGAR* (Léget et al. 2020)
- *ParSNIP* (Boone 2021)

- *SNooPy* https://csp.obs.carnegiescience.edu/data/snpy (Burns et al. 2010)
- *BayeSED* https://bayesed.readthedocs.io/en/v2.0/index.html# (Han & Han 2012, 2014, 2018)
- *EzTao* https://github.com/ywx649999311/EzTao (Yu & Richards 2022)

C.4.2. *Tools to Facilitate Anomaly Detection and Characterization in LSST Time-Series Data*

Contributors: Alex Gagliano (gaglian2@illinois.edu), W. Garrett Levine, Neven Caplar, Ashish Mahabal, Catarina S. Alves, Matthew Graham, Andrew Bradshaw, Andrew Connolly, David Trilling, Fabio Ragosta, Tomas Ahumada, Jing Lu, Ilija Medan

C.4.2.1. *Abstract*—Rubin will generate time series photometry for tens of billions of sources in multiple filters to unprecedented depth over its ten-year survey. These data will contain a dizzying breadth of persistent variable and non-variable phenomena, and some of these will be observed for the first time. Here we primarily concentrate on the data released through LSST alerts and the anomalies lurking in the alert stream. To optimize the use of the Rubin Observatory as a discovery machine for rare and high-priority events, infrastructure must be developed to efficiently identify anomalies among massive datasets in a timely manner. This will require synergy between state-of-the-art machine learning tools for anomaly detection and visualization techniques for interactive and low-latency high-level analysis. These proposed tools should be sufficiently scalable and fast enough to enable prioritization and follow-up of rapidly-evolving events before they dim (such as, early SN interaction, cometary outbursts or breakup, rapidly changing AGN, microlensing events/TDEs/kilonovae/other unknown phenomena and extreme cases of known types).

C.4.2.2. *Science Cases Needing this Tool*—Detecting anomalies in the alert stream will be essential for identifying:

- Comets experiencing volatile outbursts or breakup
- Intrinsically-rare known classes of transients (kilonovae/TDEs/FBOTs)
- Mapping SMBH Near Fields with Microlensing
- Connection between short term variability of AGN and their long term behavior
- Extreme cases of known types of variables

In addition to elucidating the physics powering known signals, anomalies will also represent entirely new classes of phenomena probing poorly-explored regimes in event brightness and timescale. The science enabled by this infrastructure cannot be fully known in advance but will have significant impact, ensuring the legacy of LSST as a pioneer in transient discovery for decades to come.

C.4.2.3. *Requirements for the software*—Anomalies in the event stream will manifest themselves in difference images and require low latency access. Because of the value of rapid follow-up, anomaly detection algorithms must keep pace with the alert stream and should be able to operate on as few bits of information as possible. Provided that anomaly detection is sufficiently fast and high-level processing is done by Rubin, the continuous alert stream could include a flag specifically for these events. Deriving meaningful information from anomalies will require flexible frameworks that can be re-trained and improved with rapid feedback processes, as many of the events we aim to discover will be "unknown unknowns" before Rubin comes online. Software that can accurately characterize astronomical events

in real-time remain woefully absent from the literature, and this presents a barrier to survey readiness.

The survey's sensitivity to unusual variable phenomena is limited by the template generation and subsequent image differencing conducted by the survey. Where possible, the template images and strategy for generating them should be reported, and the raw images should be provided through the data releases in case users wish to construct custom templates for their science case. The Rubin Science Platform could also allow for users to construct custom difference imaging pipelines close to the data to avoid duplication of data.

While high-redshift events will be contained to small postage stamps, active Solar System objects will require on-demand custom image cutouts (in some cases, as large as 10'x10') for full characterization. Even for stationary events, contextual information such as that provided by the host galaxy of an explosive transient is valuable for further characterizing a source at early epochs. Events that are not identified through the alert stream or for which larger cutouts are needed could be retrieved through the data releases, although this 24-hour latency would present an additional obstacle to rapid discovery. To manage bandwidth, a queue system or resource allocation framework among the scientific stakeholders could be implemented.

C.4.2.4. *Running on LSST and other Datasets*—Anomaly detection will leverage the alert stream and its associated postage stamps. Algorithms to find anomalies are expected to function more effectively with long baselines, but it would be ideal to identify anomalies in short light curves. This use case would be especially relevant for newly discovered Solar System small bodies approaching perihelia. Data releases for anomalous but persistent variable sources would be valuable for archival studies of these populations. Finally, rapid catalog cross-matching will be useful to determine the multi-wavelength properties of LSST-discovered sources and further investigate their underlying physics.

It is essential to identify the relevant software systems and train them with simulated samples in advance of LSST first light. On the hunt for extra-galactic transients spanning a broad range in progenitor physics, the data and models produced by the Photometric LSST Astronomical Time-Series Classification Challenge (PLAsTiCC) challenge have been state-of-the-art. These models have been improved (and the data made more realistic) in the Extended LSST Astronomical Time-Series Classification Challenge (ELAsTiCC), but data are still limited to the classes for which realistic rest-frame SED models are available.

C.4.2.5. *Existing Tools*—Central to the question of anomalous behavior is the "distance" between an event in question and a larger population. An event with a large distance from the population is considered anomalous. These distance measures could be calculated in data space (comparing light curves between events), parametric feature space (characterizing e.g., the period and amplitude of each light curve and comparing these), or non-parametric feature space (reduced-dimensionality embeddings of the original data). To generate feature representations of ingested data, methods often require close integration with feature extraction (period-finding) and dimensionality reduction tools (tSNE, UMAP, Principal

Component Analysis (PCA)). These tools can also be used to provide high-level summaries of large datasets, with which the astronomer may be able to identify anomalies by eye.

Methods differ in their focus on characterizing anomalies directly, or indirectly by constraining the properties of the "normal" larger population. The latter is more common, although novel methods are in active development. Isolation forests are an example of direct anomaly identification, and they are used by SNAD (Malanchev et al. 2021) to identify unusual supernovae and Sánchez-Sáez et al. (2021) to find unusual AGN (in the AGN case, a variational autoencoder is first used to generate a set of non-parametric features). A wealth of machine learning tools, however, focus on the bulk properties of the population (which can inform objects at the boundaries of classification), and the vast majority of these rely on decision trees (including boosted decision trees and random forests) for classifying common events. An event with low classification probability reported by these methods may be an anomaly.

A non-exhaustive list of open-source codes for photometric classification of supernovae include:

- *Avocado:* A gradient-boosted decision tree classifier for anomaly detection (Boone 2019).
- *ParSNIP:* A deep neural network that calculates latent parameters of light curves with an encoder (Boone 2021).
- *RAPID:* A recurrent neural network for real-time classification (Muthukrishna et al. 2019).
- *SuperRAENN:* A recurrent autoencoder coupled to a random forest algorithm for classification (Villar et al. 2020).
- *snmachine:* A framework for light curve feature extraction and classification using multiple techniques (Lochner et al. 2016; Alves et al. 2022).
- *superNNova:* A Bayesian recurrent neural network for real-time classification (Möller & de Boissière 2019).

Density-based methods are also in development (e.g., DBscan[14]; Local outlier factor[15]).

Because anomalies are defined relative to a population, there must also exist an interface near the data (ideally embedded within the Rubin Science Platform) for rapidly interacting with a large sample of alerts. A gap exists in current tools for early/real-time classification, and this needs to be resolved prior to LSST first light as it is central to enabling follow-up before an anomaly has ended. Paths to resolving major current barriers include:

- Introducing a broad diversity of unknown signals, parameterized by phenomenology, into simulated data spanning multiple orders of magnitudes in brightness and timescale.

[14] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html
[15] https://scikit-learn.org/stable/auto_examples/neighbors/plot_lof_outlier_detection.html

- Training and validating softwares that can rapidly recover these signals from simulated data streams with noisy and incomplete data (a few e.g., <5 photometric points and small postage stamps).
- Building tools that are highly scalable to millions of objects with runtimes of ~minutes or less
- Constructing software that is adaptable to new datasets and rapid re-training (active learning)

In the direction of flexible detection methods, Astronomaly (Lochner & Bassett 2021) has recently proposed a general framework in which active learning is used to leverage user-identified outliers and inform automated anomaly detection. Active learning in astronomy remains underdeveloped, but will likely play a powerful role in facilitating rapid anomaly identification and follow-up in the first few years of LSST.

C.4.2.6. *Computational Workflow* —The raw alert stream data (postage stamps and immediate photometry) will be accessed through community alert brokers.The brokers and/or Rubin Science Platform should provide well-calibrated historical light curves (including non-detections via forced photometry) across all diaSources at the location of the alert. Photometry for Solar System objects, which will be moving relative to the background stars, will need to track the objects themselves.

Real-time anomaly detection with LSST will likely be conducted as follows. Anomaly detection algorithms should first rapidly analyze newly acquired photometry within the context of prior light curve data. Global features such as period, parametric fit parameters, and dimensionality-reduction components should then be calculated for each event. Local features computed across windowed components of a light curve (e.g., outbursts, changing states) should also be computed to help identify anomalous behavior within an otherwise common event.

In addition, anomaly identification routines must qualitatively and quantitatively estimate the similarity between global and local features of single events and for these events in comparison to all events across the LSST catalog. This could occur via machine-learning tools to flag high-distance objects/phenomena (the quantitative approach), or visualization tools for assessing the spread between these scales (the qualitative approach). Ideally, this work will be done in parallel.

Finally, the anomaly detection infrastructure should be integrated into the LSST monitoring routines for individual working groups. These groups will need to query enlarged postage stamps and acquire high cadence Target of Opportunity photometry for extremely high-value objects. Supernovae will also need to be classified via spectroscopic follow-up.

C.4.2.7. *References for Further Reading* —

- Muthukrishna et al. (2019)
- Aleo et al. (2022)
- Lochner & Bassett (2021)

- Malanchev et al. (2021)
- Martínez-Galarza et al. (2021)

C.4.3. *Feature extraction and statistical representation (including period finding)*

Contributors: Alekzander Kosakowski (alekzander.kosakowski@ttu.edu), Fabio Ragosta (fabio.ragosta@inaf.it), David Trilling (david.trilling@nau.edu), Matthew Graham, Neven Caplar, Ashish Mahabal, Mark Popinchalk, Juan Luna, Tomislav Jurkic, Andy Connolly, Viviana Acquaviva, Catarina S. Alves, Tomas Ahumada, Jing Lu, Ilija Medan, Garrett Levine

Summary:
Many well-built and useful tools are available, but running many of them efficiently is difficult. We considered the design of a decision tree algorithm to efficiently classify variables and transients in (nearly) real-time. The full algorithm only runs on objects that produce a specific number of alerts and deeper analyses are only triggered based on probabilistic outputs from previous steps.

C.4.3.1. *Abstract* —Variable sky astrophysics is a vast field requiring many period-finding and feature-identifying tools. Thus, to efficiently identify all sources of variability in (nearly) real-time for classification and catalog-creation, a complex multistage algorithm is required. Currently, users must make use of many other tools to handle these different types of variability, resulting in running many similar analyses on the same data set to tease out different features (see Ivezić et al. (2014); VanderPlas (2018); Kennamer et al. (2020), and references therein).

Here we propose to create a hierarchical classification algorithm designed to handle the entirety of the variable-sky database generated by LSST, in combination with other southern-sky surveys. The goal of this tool is to quickly and automatically classify transients and variables based on features in a multi-band light curve (shape, period, filter-specific amplitude and decay, etc) and the available color, magnitude, parallax, angular diameter information from LSST. The proposed tool will handle multiple forms of variability simultaneously (eclipses vs pulsations vs non-periodic) in an efficient manner. Classification based on the light curve and LSST-specific photometric/astrometric data will be built-in and presented in the output as probabilities of each object being a specific type of variable.

The proposed algorithm will trigger advanced data analysis on objects that generate more than N alerts across multiple surveys (requiring crosstalk with multiple brokers). Based on the type of alert relative to the rest of the object's data, a specific set of algorithms would be triggered to identify potential periods or the presence of features to within some probability. Further analysis would be limited to cases that show specific probabilities greater than some threshold to reduce overall computation time and prevent running unrelated analyses on all data.

C.4.3.2. *Science Cases Needing this Tool* —All variable-sky astrophysics science cases will make use of this tool for initial identification leading to targeted follow-up. This tool aims to provide a first guess – which will be quantified as a probability to be labeled as a certain phenomena – on the classification of the event based on the analysis of the extracted features.

List of science cases include:

- Asteroid lightcurves
- Eclipsing binaries
- Interacting binaries
- Novae
- Kilonovae
- Supernovae
- AGN
- TDE
- Stochastic transients (e.g. stellar activity)

C.4.3.3. *Requirements for the software* —In general the input data will be a catalog photometry (in multiple bands). The algorithm would need to simultaneously access the calibrated light curve data and individual stellar color and astrometry information. Connections with Gaia Data Release 3 (DR3) will be useful until LSST can provide its own precise parallax measurements. The advanced analyses will be run on set times with increasing time-gaps between runs. (Month1, Month2, Month4, Month6, Month8, Month12, etc)

There may be cases where alert data is needed (e.g. follow up of anomalous transients). Thus, a well defined data model will give the opportunity to consider a common hyper-plane to compare and discriminate between different transients.

A single multi-column output table will be saved for each object identified in LSST. The columns can be simple summary statistics using 32-bit floats or Booleans to represent the data in as little space as possible. Intermediate steps used for creating this table will be deleted. Since different science cases may have very different feature needs, it is difficult to write a general description of the exact features will be required. The user will work with the final table for generating their own user-specific output visualization using Astronomical Data Query Language (ADQL) or some other search function.

Core requirements of the tool include:

- Ability to scale to high volumes of data. In principle, nearly every LSST source will need to be searched for periodic (or time-varying signatures). The algorithm will be expensive to run.
- Approaches need to be able to produce results for irregularly spaced data and/or data gaps.
- LSST will need to communicate and crossmatch with other surveys, including non-public databases uploaded by the user to their collaboration's RSP server.
- To accurately calculate periods for periodic variables over many years, precise timing measurements must be used.

Additional difficulties include:

- Assigning proper weights to each subclass of variable or transient is a difficult task. If the thresholds are too strict, then the resulting table is borderline useless, missing many real detections. If the thresholds are too relaxed, then the runtime will increase drastically and the output table will be filled with false positives.

- We need access to a large dataset able to provide estimates to these weights for all possible class (and potentially subclass) or variable or transient.

C.4.3.4. *Running on LSST and other Datasets* —Some LSST science cases will be adequately served by the standard Lomb-Scargle periodogram approach. Examples include, but are not limited to, variable stars and asteroid lightcurves, though we note that there are special problems associated with moving targets.

Some LSST science cases will NOT be adequately served by Lomb-Scargle and will require different approaches. These include eclipsing systems like stellar binaries or irregularly varying astrophysical sources Naul et al. (e.g. AGN, "Boyajian's Star", TDE, see B.6.9 for other examples and for examples of analysis on unevenly spaced data in time domain see 2018, and reference therein).

C.4.3.5. *Existing Tools* —There are many existing tools available for period-finding and feature-extraction. A short list of existing tools that this group is aware of:

- Celerite Foreman-Mackey et al. (2017): python library for fast and scalable Gaussian Process regression in one dimension;
- Astropy/scipy/Sci-kit learnGrisel et al. (2022) : python packages that allow users to access open source projects for specific scientific needs;
- FastAPI[16]: a modern, high performance, web framework for building APIs;
- Spark[17] is a unified analytics engine for large-scale data processing.
- Dask Rocklin (2015) is a flexible library for parallel computing in Python.

All these tools are used for a very diverse environment of transients and variable characterization in a very specific way, due to the fact that each of these tools need its own data model to be ingested to work on the reference science case. The main downside of what has been described above is the lack of homogeneity in methods for feature extraction. Nevertheless the impressive amount of data we will need to handle with LSST will not easily be tackled with this ensemble of tools because they are not obviously scalable.

C.4.3.6. *Computational Workflow* —Because of the extreme computational requirements of running many period finding approaches on every potentially varying source, one challenge will be to identify when there is a variable source that is not being well characterized by Lomb-Scargle. That is, before deploying additional algorithms, we need to identify sources that are (i) varying and (ii) not sinusoidal. How to do this? What is the scale (number) of sources that fall into this category? This could probably be estimated (i.e., how many eclipsing binaries will be in the LSST catalog, how many multi-periodic systems will be identified, etc.). This would then allow an estimate of the amount of compute resources needed to carry out this "second level" processing.

Workflow requirments include:

[16] https://fastapi.tiangolo.com/
[17] https://github.com/apache/spark

- Access the entire LSST data archive for light curves, colors, parallaxes, etc
- Access to the alert brokers would be needed to provide intel to follow ups, the software would be run on sources detected in the alert n (we considered n< 4) times. The cross-talk between the software and the brokers will be imperative for early classifications.
- Pegasus enables scientists to construct workflows in abstract terms without worrying about the details of the underlying execution environment or the particulars of the low-level specifications required by the middleware (Condor, Globus, or Amazon Amazon Elastic Compute Cloud (EC2)). Pegasus also bridges the current Cyber Infrastructure by effectively coordinating multiple distributed resources.
- The output will be a simple multi-column table. Preferably with classification columns appended to the standard color/magnitude/parallax columns to save disk space.

C.4.3.7. *References for Further Reading* —Richards et al. (2011), Narayan et al. (2018), Muthukrishna et al. (2019), Forster et al. (2020)

C.5. *Image Reprocessing*

Contributors: Francois Lanusse, Joachim Moeyens, Steven Stetzler, Suvi Gezari, Clare Saunders, Matt O'Dowd, Charlotte Olsen, Alma Gonzalez, Gabriele Riccio, William O'Mullane

C.5.1. *Abstract*

There will be a need for reprocessing of images or image cutouts for a variety of science cases as listed below. In some cases this is making custom cut outs for specific types of objects, potentially reprocessing the cutouts and stacking them in a new coadded image. This is distinct from reprocessing full images ala data release processing. In some cases custom processing of the single frame cutouts is requested, this may require changing the background subtraction, extraction of photometry,

C.5.2. *Science Cases Needing this Tool*

Cosmology:

- Strong lensing scene modeling, SN scene modeling, (potentially) Cluster lensing specific deblending and shape measurement pipeline

Extragalactic (Static):

- SED fitting (Section B.2.4): LSST's extensive deep imaging of the southern sky will allow for synergies with multiple other legacy surveys which have different wavelength coverage. With six bands, the amount of galaxy properties which can be recovered through SED fitting is limited to photometric redshifts, colors, and stellar mass. Adding additional supplementary IR bands from surveys such as VISTA, Spitzer, and WISE will allow us to constrain more galaxy properties as well as reconstruct star formation histories with sufficient bands. For this reason it will be necessary to reprocess LSST images to the same PSF of images from crossmatched galaxy catalogs. There is a need to develop and/or implement algorithms that will allow images to be convolved to match pixel scales of IR sources. Some of this work is being done and tested with HSC, but making a user friendly interface will greatly serve the community. It should require little when using on the fly calculations, and the only data products that may be necessary for LSST to store would be catalogs of fluxes for crossmatched galaxies from various other surveys.
- Low surface brightness dwarf galaxy (candidate) catalogs out to 100 Mpc (Section B.2.1). A good fraction of nearby dwarf galaxies are very low surface brightness objects, on which the source extraction algorithm may not perform well. At the image level one would need to obtain calibrated cutouts for this sample together with flags from the deblending process and re-fit the photometry with models that are optimized for LSB dwarf galaxies. This would produce a new catalog with matched photometry and colors.
- Lens discovery where we would like to identify the optimal image coadd if different from the default pipeline coadd.

Extragalactic (Transient Science):

- Deep custom stacking of pre-event and post-event epochs to search for pre-cursor eruptions and measure late-time evolution, and to create weekly/monthly stacks and do difference imaging to probe faint and slow transients
- Deep custom stacks of AGN in a high state vs low-state, to isolate the position of the AGN relative to the host galaxy nucleus wandering/recoiling MBHs, IMBHs in dwarf galaxies (Section B.4.6)
- Custom deblending of small-separation gravitationally lensed quasars for lens finding (Section B.4.7), microlensing studies, dark matter substructure, and time-delay cosmography

Solar System:

- Active Objects I/II (Section B.7.2): Determining the presence of activity (sublimation, outgassing, collisional breakups) of Solar System small body populations requires processing custom-sized and custom-shaped nightly cutouts to look for evidence of activity. This would require querying for, at times, greater than >O(100x100) pixel cutouts of difference images for each observed Solar System object (100k-100m in a night). Active object cutouts should also support custom-stacking for pairs of nightly observations, and on the larger-scale, support stacking over ~month-long cadences. This would include specifying orientation such that any tails and/or motion can be aligned in the stacked images.
- Searching for Faint Objects (Section B.7.6): There is a large population of slow moving solar system objects (KBO/TNO/Planet 9) that LSST will observe below the 5 sigma source detection limit of a single exposure. Objects from these faint populations can be recovered using efficient shift-and-stack algorithms accelerated with GPUs (KBMOD) on difference images produced by LSST. This use-case requires access to either user-defined sized cutouts of difference images or access to individual visit focal planes if user-defined cutouts cannot be supplied. To validate moving object discoveries in the difference images, we would perform forced photometry on cutouts of the calibrated exposures centered on the discovered objects locations. A shift-and-stack search that is both deep and complete would require access to images across the ecliptic (~3600 square degrees) and across a full year of LSST data, with these searches being run quarterly or yearly. These searches would significantly benefit from having public/group access to the images at a supercomputing center or in the cloud.

C.5.3. *Requirements for the software*

The given science cases require a few different areas of functionality:

- Several science use cases will require custom-sized and custom-shaped cutouts of both calibrated exposures and difference images, and these will also be required for some of the functions described below. This can be provided by an image cutout

service that allows users to access cutouts of a chosen size for visit and coadd-level images.

- Extragalactic transients and variable objects required custom image stacking. This would most likely be able to use the existing coaddition algorithms, but would require custom combinations of visits into the coadds. For example, transients would benefit from stacking all the images before or after the lifetime of the transient, or AGN images could be separated into bright and faint time periods.
  - Existing tools: the existing coaddition algorithms in the LSST data processing can be used to do this. Also, Zuds is a tool from ZTF that performs custom stacking.
  - Needed tools: Custom stacking could be made more simple for users with a tool that would allow date ranges to be specified for the input images.
- Alternatively, some nearby objects will require more specialized stacking: to find nearby Solar System objects, images need to be shifted and stacked following the likely trajectory of an object. An additional allowance to specify the orientation of each image in the stack is desired so that the tails of active objects can be aligned in the stack.
  - Existing tools: KBMOD: GPU-based shift-and-stack code to search for Solar System objects
  - Needed tools: KBMOD still needs significant development and infrastructure support to work at the scale of LSST, custom-sized cutouts of difference and calibrated exposures
- Custom deblending will be necessary for analyzing gravitational lenses, where software specific to lensing must be used. Other particular use cases such as low-brightness dwarf galaxies will want to spot-check the standard deblending.
  - Existing tools: scarlet, BEAST, GaaP , lens modeling tools (e.g. lensStronomy, molets, Muscadet)
- Custom photometry will be required for science use cases where it is necessary to combine LSST measurements with external datasets by reconvolving the data.
- Scene modeling on the calibrated visits will be required for Type Ia Supernova cosmology and for strong lensing.
  - Existing tools: Tractor and scarlet, others tailored to specific projects
  - Needed tools: Something that takes advantage of existing LSST data products, such as the per-visit calibration and WCS information

C.5.4. *Running on LSST and other Datasets*

This service would run on any image cut outs from the Rubin image cutout service. In principle this could be run on any available multiepoch image dataset with a cut out service.

C.5.5. *Computational Workflow*

There may be more than one use case here but all start with image cutout of some sort. The image cutout service should be a good starting point.

1. Select an object or set of objects make appropriate sized cutouts - these may be long polygons for solar system objects
2. For each cutout apply custom processing if required (reconvolve for matched catalog)
3. For each object stack the images
   (a) If this is time batched make the stacks according to the given time frames (implies perhaps a single object in step 1)
   (b) If this is a shifted stack (epoch centered) the images should already be centered on the object and can be stacked ignoring positon.
   (c) Otherwise standard stacking/dithering should be applied.

Shifted stack:

1. For a patch of sky get cutouts from the difference images at all epochs
2. Pass this set of images ot the custom GPU code for stack shifting
3. For candidate objects go back and get cut outs and forced photometry from the locations provided.

C.5.6. *References for Further Reading*

Functionality Required for the technical Case:

- Custom coadds (for extragalactic transients), using different time periods (i.e. before and after lifetime of transient; bright and faint times of AGN) - these would use custom times, but not custom stacking algorithms
- cross-matching : reconvolving LSST photometry to combine it with photometry from external datasets (co-processing) would be needed for coadds and at visit level
- Gravitationally lensed quasars: want to do own deblending, some may be urgent, so need to be done in real-time.
- Solar System active objects: look at the cutouts and see if there is an extended body for all Solar System objects observed in a night. May need >100x100 pixel cutouts. Pull out cutouts over the history of a moving object.
- Solar System faint objects- custom coadds: shift and stack along the trajectory of Kuiper belt objects - can get to 26th magnitude using all images from first year. Might need arbitrary polygons. Searches are done using GPUs
- LSB dwarf galaxy catalog out to 100 Mpc: want to repeat photometry using models that are optimized for dwarf galaxies. Want the deblending information
- Lens discovery - are coadded images optimal for lens discovery? May need custom coadds in order to see whether regular coadds are sufficient.
- Type1a supernovae - reprocess with scene-modeling get highly calibrated measurement for point source.

- Static local universe - go back to images to validate whether stars and galaxies have been accurately distinguished.

Existing Software Needed:

- Image cutout service (which is already provided by RSP)
- 100K to 1M cutouts per night - bigger than alert cutout since they are extended (solar system)
- There are scene-modeling codes - Tractor, Scarlet,
- KBMOD for Solar System shift and stack
- Deblending tools for lens modeling (lenstronomy, Muscadet)
- BEAST, GaaP (coming soon to DM)
- Need algorithm for reprocess LSST images to match the pixel scale/PSF of complementary IR surveys (e.g., Tractor)
- Need a means for users to query other catalogs to pull matched images/IDs for crossmatched galaxies
- Zuds - custom stacks for transients (ZTF)
- Monthly and weekly stacks for long lasting fainter transients
- Galaxy morphology fitting (e.g., GalFit)

New data products:

- New fluxes for SN lightcurves – on the order of 1M data points
- 10k per year SN-1a - 100K objects, postage stamps and possibly new coadds
- 8k lensed quasars need light curves; ~200K lensed galaxies need cutout images
- 100K- 1M Solar System cutouts per night (custom shapes)
- Euclid, Roman etc co-processing with LSST data to search for lenses (petabytes of data)

C.6. *Image Analysis*

Contributors: Federica Bianco (fbianco@nyu.edu), James Chan (hung-hsu.chan@epfl.ch), Colin Orion Chandler (orion@nau.edu), Henry Hsieh (hh-sieh@psi.edu), Arun Kannawadi (arunkannawadi@astro.princeton.edu), Ilin Lazar (i.lazar@herts.ac.uk), Yao-Yuan Mao (yymao.astro@gmail.com), Knut Olsen (knut.olsen@noirlab.edu (facilitator)), Tyler A Pritchard (TylerAPritchard@gmail.com), J. Antonio Vazquez-Mata (jvazquez@astro.unam.mx)

C.6.1. *Abstract*

While much of the science that will be done with LSST will rely entirely on the catalogs delivered by the science pipelines, a significant number of science use cases will also require analysis of the image data. In this section, we summarize the discussions of the kinds of software tools that are likely needed to enable image-based analyses of a broad range of science use cases. In this breakout, we only considered analysis that would be done at the level of objects; large-scale image analysis is covered in the section on image reprocessing.

We first briefly summarize the ways in which the science use cases will use the image data, then enumerate the software capabilities required by one or more of these use cases.

C.6.2. *Science Cases Needing this Tool*

- Lensed quasars: Section B.4.7
- Galaxy morphology: Section B.2.3
- Transient host galaxies: Section B.3.1
- Slow transients: The Rubin Observatory LSST Alert pipeline's difference imaging pipeline will efficiently capture transient and variable events with rapid rises in brightness. However, one weakness to this methodology is the reliance on templates that may contain flux from sources that vary slowly over timescales comparable to that between template image acquisition. Slowly evolving events such as high-redshift ($z > 2$) SLSNe and Type IIn supernovae with timescales of a year or more can be difficult to detect in these classical nightly image subtraction surveys with fixed templates, and can require special processing to detect including image subtraction with custom templates or bespoke ML-algorithms. Custom image analysis on galaxy cutouts with the tools described below will allow for the expansion of our ability to detect these events and enable them to be discovered more promptly and enabling follow-up studies.
- Detection and analysis of active small solar system objects, Section B.7.2: The detection and analysis of visible mass loss from comets and asteroids due to various processes (e.g., sublimation, impacts, rotational disruptions) are a high priority for LSST solar system science. This science case includes the detection of previously unknown activity, characterization of morphology evolution of known active objects, and the detection of anomalous brightening of known active objects (i.e, cometary outbursts).

- Distances for < 100 Mpc dwarf galaxies Section B.3.1: Use ML-based methods (e.g., CNN) to estimate distance for very nearby (< 100 Mpc), low surface brightness dwarf galaxies. The algorithm will run on postage stamps, with a small training data set based on a subset of objects with known distances/redshifts. Existing photo-z algorithms are not optimized for this very nearby region, and the lack of training data also poses a special challenge.
- Weak gravitational Lens cosmic shear (WL) artifacts/debugging: Failures and outliers in the measured shape of seemingly normal sources may be caused due to a neighboring source, which can be easily identified from image cutouts. Examples of this include: i) a bright star in the vicinity, elevating the local background ii) an extended arm of a spiral galaxy or merger components of an irregular galaxy with a dominant bulge component (so that it is not evident from the catalogs) overlapping a faint source of interest. A quick look at the image can differentiate an algorithmic failure of the measurement from imperfect deblending.
- Search for resolved dwarf galaxies, Section B.5.2: A complete census of dwarf galaxies within the Local Volume will rely on both catalog analysis of resolved stars and analysis of the associated images (e.g. Carlin et al. 2021). The image cutouts will need to be up to ~10 arcminute minute of arc (unit of angle) (arcmin) in size to contain the candidates and their context. The goal of the search will be to remove humans from the loop to the extent possible, and/or use the citizen science community to help.

C.6.3. *Requirements for the software*

Because image-based analysis is a very broad technical subject, we first listed the software capabilities needed, identified the science cases needing each capability, and defined the data products that each capability would need as input to the analysis. Later in this section, we will describe the requirements of the capabilities in more detail.

- **Capability:** Image cutouts. While also discussed in the Image Reprocessing breakout, all image analyses discussed presume the ability to make image cutouts of any input image over scales ranging from ~10 arcsec to ~10 arcmin.
  Needed by: All.
  Input data products: Coadds, single epoch images.

- **Capability:** Ability to perform custom deblending and reblending of the objects detected in an image.
  Needed by: Lensed quasars, transient host galaxies, distances to dwarf galaxies within 100 Mpc, weak lensing artifacts and debugging.
  Input data products: Coadds, clipped coadds.

- **Capability:** Ability to link an image cutout to archival data (both catalogs and images) from external sources.

Needed by: All.
Input data products: Images, catalogs.

- **Capability:** Ability to conduct analyses on any band or combination thereof.
  Needed by: All except the comet activity use case.
  Input data products: Coadds, single epoch images.

- **Capability:** Ability to conduct image analysis in real time.
  Needed by: Transient host galaxies, comet activity.
  Input data products: Single-epoch images.

- **Capability:** Ability to associate image cutouts with measured light curves.
  Needed by: Lensed quasars, transient host galaxies, slow transients.
  Input data products: Coadds, single-epoch images, catalogs, alert stream.

- **Capability:** AI-aided image analysis and time variable clustering analysis. The idea behind this capability is to use machine learning and AI techniques to help in the construction of difference images, their analysis, or to replace traditional DIA entirely.
  Needed by: Lensed quasars, slow transients.
  Input data products: Coadds, single-epoch images.

- **Capability:** Machine learning-based clustering of static images.
  Needed by: Galaxy morphology, comet activity, distances to dwarf galaxies within 100 Mpc, search for resolved dwarf galaxies.
  Input data products: Coadds.

- **Capability:** Ability to handle diverse morphology found in images.
  Needed by: Galaxy morphology, comet activity, distances to dwarf galaxies within 100 Mpc, search for resolved dwarf galaxies.
  Input data products: Coadds, single-epoch images.

- **Capability:** Ability to conduct visual inspection of images.
  Needed by: Galaxy morphology, comet activity, distances to dwarf galaxies within 100 Mpc, weak lensing artifacts and debugging, search for resolved dwarf galaxies.
  Input data products: Coadds, single-epoch images.

- **Capability:** Ability to use images in citizen science programs, which is being delivered as part of Rubin Construction (and thus not discussed further).
  Needed by: Galaxy morphology, comet activity, distances to dwarf galaxies within 100 Mpc, weak lensing artifacts and debugging, search for resolved dwarf galaxies.

Input data products: Coadds, single-epoch images.

- **Capability:** Ability to create synthetic image cutouts or inject objects into empirical image cutouts.
  Needed by: All.
  Input data products: Coadds, single-epoch images.

C.6.3.1. *Image cutouts*—The image cutout tool that returns postage stamp(s) must be able to take in as input at least one of the following:

- Positions (RA, Declination (DEC)) on the sky and a size of the image, either in arcsec or in number of pixels. Depending on the science use case, we expect the size of the postage stamps to vary from 10 arcsec x 10 arcsec to 1 deg x 1 deg
- Unique identifier of a source/object from the Data Release catalogs and return postage stamps with size determined by the footprint, along with a buffer optionally.
- The postage stamps should contain sufficient metadata information to produce color-composite images from stamps that will identify objects such as lensed quasars.
- Single-visit images should be queriable by metadata information, e.g., observations of (RA, DEC) between 'datetime1' and 'datetime2' or seeing size < 0.5 arcsec, etc. This can be used to look for transient/variable objects specifically. (new feature?)

The returned stamps can be single visit images, or any of the coadds produced by DM. The returned images must be viewable on a portal on the browser for some quick visual inspection and allow for bulk downloads, for creating ML datasets. The portal must allow for an interactive visualization, i.e., zooming in, finding pixel values, variance estimates and mask bits for each pixel etc.

Existing tools, such as that provided by the RSP, have most of the above capabilities, but some work may be needed for high-level capabilities such as integrating the cutout service with Table Access Protocol (TAP) query that can translate ADQL queries to image cutouts through the catalogs.

C.6.3.2. *Deblending / reblending*—High-level requirements for the deblending/rebending tool are:

- Being able to specify postage stamps to run on.
- Being able to run on both coadd postage stamps and clipped coadds.
- Being able to customize the deblending criteria. Should have a well-defined Application Programming Interface (API) so that users can supply their custom deblending criteria / algorithms easily.
- Being able to customize the photometry fitting procedure. Users can customize how to refit the photometry and produce a new catalog (for the all postage stamps that it has run on).
- Generate some sort of scores based on user-supplied metrics.

- This will mostly be used for static science, so it will not be run very often. It might need to be run on a large number of postage stamps, but that should still be a small fraction of the overall sky coverage (e.g., < 2%). For alternative deblending/reblending methods that need to be run on a large fraction of sky, it should be part of the image reprocessing.
- The data output (catalog) needs to be stored for a long term for science use.

Existing tools:

- Scarlet (Melchior et al. 2018), Source Extractor (Bertin & Arnouts 2010), SDSS Deblender, Tractor (Lang et al. 2016), GalFit.

C.6.3.3. *Link to archival data*—For external archive data to be used in conjunction with LSST data, the science use cases require:

- Object-based and coordinate based external archive queries
  - Moving objects:
    - something like the CADC SSOIS[18] or PDS CATCH[19]
    - Faster yet: already linked to objects (e.g., ZTF alert stream)
  - Maintain tables of crossmatches and archive-specific metadata in order to facilitate productive external archive searches
  - Augment table data with potentially missing metadata (e.g., depth, seeing, . . .)
- Data products would include images and catalog objects
- Programmatic access important
- Simple Image Access (SIA) interface for images
- TAP interface for catalogs
- Build on Astroquery as a general interface
- Local or Cloud storage for data products
- Ability to store products in multiple file formats

C.6.3.4. *Real-time analysis*—This is a general software requirement/consideration rather than a specific software tool. Example science use cases that require Real Time Analysis (RTA) include the following:

- Detection of active solar system objects (RTA needed to enable follow-up
- Comet outburst detection (RTA needed to enable follow-up)
- Characterization of transient host galaxies

Requirements for software performing real-time analysis include the following:

- Automation of image analysis tasks for specific science cases (e.g., active solar system object detection, comet outburst detection, characterization of active solar system object morphology, analysis of transient host galaxies)

[18] https://www.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/en/ssois/
[19] https://catch.astro.umd.edu/

- Rapid definition of postage stamp requirements (e.g., exposure, position, cutout size)
- Prompt retrieval of postage stamps from single-epoch data via image cutout tool
- Same-day completion of data analysis tasks for all data acquired in a single night at a minimum given the need to keep pace with overall data acquisition rate
- For some applications, near-real-time analysis may be desirable to enable extremely rapid follow-up for highly variable targets
- Job management infrastructure to manage massively parallel data processing to achieve required processing speeds

C.6.3.5. *ML-based "DIA" / variable clustering —*

- DIA is an expensive and critical step for Time Domain Astronomy. DIA models will be tested throughout Rubin commissioning to select the most effective method. Critical with Rubin will be the need to limiting false positives and accurately perform Real/Bogus due to the enormous survey data volume (expected 10M alerts per night).
- Traditional DIA will rely on the construction of a template which Rubin will update annually (with each data release). All LSST historical data will be reprocessed with new template in each data release. DIA needs to be efficient, with the current computational bottle neck being the PSF matching operations that align the properties of the images that compose the template and those of the template with those of each science image.
- We expect that a traditional DIA analysis will be the basis for the alert generation throughout the survey lifetime. A typical DIA workflow entails Template creation, PSF matching, Image Subtraction, Transient Detection-Real Bogus (typically done with feature based methods like Random Forests), Photometry, Transient classification (typically performed on multiple data points). Yet we expect that the community will want to customize the transient discovery process to, for example, increase sensitivity to specific classes of transients by designing purpose-specific templates and/or modifying the traditional DIA workflow. This workflow can be modified in several AI-aided ways. An AI model could be developed for progressively more complex tasks thus replacing more of the transient detection and characterization infrastructure. A simple minimal task would include AI-aided detection and/or real bogus which requires a binary classifier (a many to one NN classification), more complex tasks could include transient classification (a many to many NN classification) or photometry (regression), or any combination of the tasks outlined here. Example tasks include:
  - Comparing the science image with a historical (LSST but potentially also precursor surveys) collection of images from the sky position (i.e. the input of a neural network would be a data cube of historical images and a single science image, the output would be a transient detection/real bogus classification/transient class classification). This is expected in particular to increase effectiveness in the detection of slow evolving transients where an average template

  - AI-generation of templates: this entails teaching a neural network to perform all the computationally expensive steps of template creation including warping, normalizing, and PSF matching of the template and science images
  - Bypassing the PSF matching and image subtraction steps by comparing the template with the science image (steps in this direction have demonstrated the feasibility of this approach (Acero-Cuellar et al. 2022) in the Real Bogus step, but the detection step and transient classification without DIA are yet to be explored. . .).
  - Direct transient classification from DIA images, template+DIA images (providing critical context for the transient's origin), template + science image alone, or a collection of images from the same sky position (no template), which is currently being explored (e.g. ALeRCE[20]).
- Need for efficient image subtraction with a variety of templates or template-agnostic image subtraction
  - For example, slowly evolving objects (High-Z SLSNe, IIn), sub-threshold events (High-Z transients), variable events with data-in-templates (AGN), non-point source/diffuse emission (e.g. light echoes), periodic and semi-periodic variabilities with time scales that are on times scales comparable with the cadence of images selected for templates (Hambleton et al. 2020)
- Requirements
  - Depending on methodology, AI-aided DIA can use a template image and science image to most closely mirror today's DIA workflow, or a time-series of single visit images to search for events
  - Robust access to multi-band single visit images as well as yearly/deep stacks depending on choice of AI-aided methodology - postage stamps may be sufficient for this so long as they are large enough to enable matching (large enough to characterize the PSF) or PSF modeling information is provided
    * One may choose to do this on a variety of scales depending on the science use case - Deep-Drilling Fields for exceptionally long-lived events, a known galaxy list (that can be exceptionally large, e.g. 2<z<5) for subthreshold events, the entire field for diffuse/light-echo type emission
  - Efficient access to ML-tools, computational cores, and rapid retrieval on the terabyte up to approaching petabyte scale data
  - Significant additional storage for synthetic or historical training data
  - Ability to re-process as tools improve
- Challenges
  - Prototyping of many models similar to this methodology exist in the literature, but have never been run at anything approaching scale. To do this, we need:

[20] http://alerce.science/

* Training - for a truly robust AI-aided DIA algorithm one must have sufficient training data to capture the known on-sky distribution of objects. This will require simulation, precursor, and on-sky data to work in concert.
* Data-scaling - at the largest scales, this requires robust processing across the entire sky of data. While the simplest examples of this are trivial (e.g. a small number of point sources from a known list of locations) it quickly scales to terabytes of data volume and velocity

C.6.3.6. *ML-based static clustering* —Morphology is a fundamental parameter, essential for the full spectrum of extra-galactic LSST science. LSST offers an unparalleled combination of depth, area and statistics, with ~20 billion galaxies expected from its 18,000 deg^2 footprint with a point-source depth of $r_{\rm AB}$~27.5 mag. A rich literature exists on measuring morphologies in surveys, from visual inspection, using systems like GZ, to automated methods, either via simple measures (Sérsic/CAS etc.) or sophisticated supervised or unsupervised machine-learning (ML) techniques. However, the unprecedented size of LSST requires a radically different approach. Visual inspection, even using GZ, will be prohibitively time-consuming. Furthermore, since the morphological detail in galaxies will increase as LSST becomes deeper, morphological catalogs will be needed at multiple depths (e.g. from every data release). This calls for classification/clustering techniques which are able to handle large amounts of high cadence survey data (petabyte to exabyte scales) in an efficient and accurate way.

- Need for supervised ML to do morphological classification of galaxies and detect particular structures, using previous labeled data.
- Need for unsupervised ML in particular for large scale galaxy morphology classification may provide an advantage since it does not need training sets (i.e., no need for labeled data)
- Photometric redshift estimation
  - Weak/Strong lensing studies
  - Studies of Galaxy evolution as a function of environment, redshift and other properties from a statistical perspective (which LSST can now enable)
- Requirements:
  - Need for combined calibrated images in all possible bands to carry out classification.
  - This can be done either combining images in the RSP during the training process or having previously combined cutouts in png format. Meanwhile the first one will require computer power to accelerate the process, the second one will require additional 2 Tb of storage for every 500 millions of 50x50 pix images.
  - Being able to access/propose for CPU time easily and efficiently (for example 100 cores for each LSST yearly release); maybe even have a portion of the LSST CPU capabilities reserved for ML based applications on a yearly basis

  - Need an efficient architecture within the RSP to be able to import training data in large scales from other surveys
  - Need for an efficient data transmission to local IDACs with GPU facilities.

  The classification process is expected to be repeated at every data release to generate morphological catalogs.
- Existing tools: External tools to develop ML algorithms have been developed and optimized to do efficient calculation. Tensorflow, Keras, PyTorch are the most accessible and friendly libraries to work with.
  The most popular technique used for galaxy classification is CNNs (e.g., Huertas-Company et al. 2015; Cheng et al. 2020b; Dai & Tong 2018) using training data mainly from the Zooniverse. Other techniques are random forest classifiers, Support Vector Machines (e.g., Goulding et al. 2018) or unsupervised algorithms (e.g., Cheng et al. 2020b). The downside is that most of these techniques require large amounts of training data and may not be able to operate efficiently at LSST scales.

C.6.3.7. *General Visual Inspection* —

- Ability to display a grid or list of relevant images.
- Ability to pull from different sources (e.g., Pan-STARRS1, SDSS)
- Mouse over information: RA, Dec, x, y, UT observing date, filter?, ..?
- Ability to flag images and export results
- Ability to change angular FOV
- Ability to rotate, flip images
- Ensure uniform orientation (N up, E left)

Examples:

- https://yymao.github.io/decals-image-list-tool/

C.6.3.8. *Testing synthetic images* —

- This is an overall infrastructure functionality, rather than a specific software.
- Being able to process user-provided synthetic images with both the LSST Science Pipeline and all the custom tools above.
- Should accept both fully simulated images and real images with synthetic source injections.
- Provide access to the produced catalogs and other data products in the same way as the real data products.
- This will mostly be used for testing, verification, and validation. It may be run more often during the development stage. However, the data products may still need to be stored for long terms as they may be needed for testing bias, completeness for science use.

C.6.3.9. *Diverse Morphologies* —

- Ability to account for a range of different possible morphologies for extended objects for various purposes
- Ability to fit user-provided models to image cutout data
- Example relevant science cases:
  - detection of different types of small solar system object activity (e.g., circularly symmetric coma, faint dust trails)
  - Recalculation of astrometry of highly active comets fitting comet-like profiles

C.6.4. *Running on LSST and other Datasets*

As discussed above, the image analysis capabilities will make use of several Rubin LSST data products, including coadds, clipped coadds, single epoch images, catalogs, and the alert stream. There also was a general need for access to external archival data. A few image analysis technical cases also had specific needs:

C.6.4.1. *ML-based "DIA" / variable clustering* —

- Existing archives of Survey Data. Time-series and static, in similar filters as the main survey
- Simulated static Sky-data from the LSST with source injection for transient and variable events
- LSST commissioning data
- On-Sky LSST Single epoch and stacked data

C.6.4.2. *ML-based static clustering)* —

- The LSST Data Previews, the HSC survey (LSST precursor), Hubble Space Telescope (HST) datasets and the DESI Legacy Surveys will be beneficial to act as training sets
- The algorithm can be run as soon as the first data release is online if unsupervised machine is used and if training is done on precursor or other data

C.6.5. *Existing Tools*

Existing tools vary by image analysis capability, but include:

- Rubin Science Platform (image cutouts)
- Scarlet, Source Extractor, SDSS Deblender, Tractor, Galfit (Deblending/reblending)
- External data archives
- Tensorflow, Keras, and PyTorch for ML applications
- The Rubin Education and Public Outreach (EPO) Citizen Science project for the citizen science use case

C.6.6. *Computational Workflows*

Detailed computational workflows were provided for the two of the image analysis capabilities.

C.6.6.1. *ML-based "DIA" / variable clustering —*

- A minimal example for this technology is developing an AI-aided detection and/or real bogus methodology for the Rubin Observatory LSST, and it could follow the following development process:
- Training for AI-Aided DIA
- Prior to LSST operations, precursor surveys and simulated data (such as that created by DESC for the Data Preview) could be used to optimize both the neural network structure (building off of extant prototypes) and training methodology
- With the release of wide-field LSST commissioning data at levels similar to both individual (and potentially 10-year) survey depth, transfer learning methodology could be used to evaluate success of the algorithms on precursor/simulated data when applied to on-sky survey-like data.
- As additional on-sky data from Rubin observatory cameras becomes available, models can be continuously improved with adaptive learning techniques to improve quality
- On-sky AI-Aided DIA
- The user would identify a list of sky-area to for the DIA - depending on the scale of the process this could be a HEALPix map, or a list of galaxy or point source positions with postage stamp size (that may come from catalog cuts or external sources)
- The user would identify a 'template image' to use, including those created by themselves, LSST simulated images from precursor surveys, or the LSST project
- The user would identify a AI-aided DIA model to use for their science project. In the optimistic case this could be a single model that has been robustly trained across the sky. In a more pessimistic case this could be something like a consensus NN infrastructure that the user trains on a subset of objects that are similar to their science needs
- The user then creates a (potentially embarrassingly parallel) processing pipeline on the RSP, LINCC, or local compute that takes template images, survey images/cutouts, and runs the AI-aided model to receive a list of detected events. This would then be a jumping off point into a further scientific study revolving around the objects of interest. . .

C.6.6.2. *ML-based static clustering —*

- We will need to query galaxy cutouts and catalog properties from the LSST archive. Example catalog properties needed: ra, dec, photometric redshift, stellar mass, SFR, colors, object radius.
- The classification/clustering will be done with algorithms provided by Scikit Learn and suitable processing parallelization will be used
- First training will be done using external catalogs and the Data Preview 0 (DP0).2 images through the RSP.

- Once the first Data Release (DR) comes out, the algorithms will be run on these images and morphological catalogs will be generated. About 1Tb of storage will be needed to save these catalogs.
- If 100 cores are used the training timescales may last for a couple of days to a week depending on the depth of the data and sky cover age.
- If GPUs at IDACs are used, the training timescale could be reduced by a factor of 2. However, It will be very important to estimate what is cheaper and more efficient, using more cores in CPUs at the RSP or moving data to IDACs with GPU facilities.

C.7. *Photometric Redshifts*

There were multiple technical cases for photo-z. The first of these is about how to effectively represent and store photometric redshift information; the other two relate to how statistical uncertainty is quantified and the impact of photometric redshift quality on scientific applications. Essential background on what Rubin Observatory will provide regarding photometric redshifts, and the anticipated connection with the scientific community in this area, is given in https://dmtn-049.lsst.io/.

C.7.1. *Photo-z p(z) representation and storage*

Contributors: Sam Schmidt, Julia Gschwend, Alex Malz, Raphael Shirley, Ashley Villar

C.7.1.1. *Abstract* —Current-generation surveys (DES, HSC, and others) now commonly supply one-dimensional (1D) photo-z Probability Density Functions (PDFs), typically stored as evalulations on a grid or as samples, though existing software tools such as qp (see reference below) are being developed to explore more efficient parameterizations of PDFs for next-generation data sets, bearing in mind the limited resources for storing and serving PDFs. However, dedicated effort will be required to minimize the inefficiencies of lossy format conversions and the computational expense of repeating estimation procedures to achieve the most appropriate format for *each* science case expected to use the PDFs. It is also important to provide documentation on how photo-z PDFs are represented and how they can be“decompressed" to a traditional grid PDF or other format, if necessary, for a particular use case.

C.7.1.2. *Science Cases Needing this Tool* —The storage of photo-z impacts all science cases where redshift is needed beyond a simple “point estimate”.

Storing a 1D PDF or a 2-dimensional $p(z, \alpha)$ (where α could be star formation rate, stellar mass, etc.) requires that decisions be made as to the format in which the data will be stored, be it on a grid, a specific set of quantile values, a mixture model fit to samples, etc. Efficient storage and the ability to actually *use* the resulting photo-z PDFs across all science cases is an essential need. That is to say, the most appropriate storage format for a PDF may be different from what a science-case-specific code has used for its input in the past, and methods should be provided to transform between representations where necessary. In some cases, adapting the metrics/analysis on the user-side to utilize the efficient storage format directly may actually improve workflow performance: for example if the Cumulative Distribution Function (CDF) is used but never the PDF, then quantiles would reduce computational expense even if past use cases assumed a gridded input format or drawing samples from the PDF. Some thought will need to go into each science case to decide on an optimal parameterization balancing storage limitations against loss of information to which downstream analysis is sensitive. In the end, the storage method must cover all science use cases.

C.7.1.3. *Requirements for the software*—Many of the requirements could be fulfilled by the qp[21] software package, which aims to be a general, extendable tool for transforming between multiple PDF representations, and includes metric computations. Multidimensional photo-z estimates, where additional quantities are jointly fit, are obviously more computationally intensive and require a more complex solution for storage. The DM PhotoZ table is expected to have ~200 columns, and even with additional database storage that might be available from the LIneA Brazil in-kind contribution, multidimensional PDFs must be able to be represented by a relatively modest number of parameters if they are to be implemented on the full LSST dataset.

For science that will involve a relatively small sub-sample (compared to ~billions of LSST detected objects, say up to millions of objects), it may be simpler and computationally less expensive to compute multidimensional distributions "on the fly" for that subset as-needed rather than attempt to store them on disk, particularly if the specific multidimensional parameterization does not lend itself to compression to a small set of parameters. While PDFs are necessary for many science cases, some still use single point estimates of redshift or low-dimensional summary statistics (e.g. moments), and thus these will also be included in any data releases; some discussion of uncertainty quantification for such point estimates is included in Section C.7.2 of this white paper.

C.7.1.4. *Running on LSST and other Datasets*—Both Rubin photo-z outputs and additional photo-z estimates are very likely to include PDF representations, as are all annual LSST release catalogs and DDFs for which photo-z's are computed. Any multiband precursor dataset could also be used to generate PDFs for experimentation.

Photo-z PDFs could be produced for any new release of photometry and any update in prior information, such as additional spectroscopy for training sets or new SED templates. The provided data products would have to include provenance information indicating the version of software as well as the version of prior information and input data used to create it. However, computational and storage expense for photo-z estimates is nontrivial, so some thought will need to go in to how often release catalogs are generated.

C.7.1.5. *Existing Tools*—qp is an existing software package to transform between several PDF representations (interpolated grid, histogram, samples, quantiles, parametric mixture model, etc.) and to evaluate metrics of 1D PDFs. qp can be used to identify the optimal parameterization for a given scientific use case by allowing users to experiment with the number of parameters and format, ensuring sufficient information is preserved in the approximation/decompression steps to achieve the desired science goals.

qp deals only with 1D PDFs currently, though there has been discussion of expanding to multidimensional, but computational and storage efficiency would be an issue. qp is still under active development, and several of the storage methods may still be somewhat slow, particularly for transformation of large datasets from one format to another, and so

[21] available at https://github.com/LSSTDESC/qp

additional code optimization may be necessary. Any extensions should be compatible with existing qp code, or entail refactoring with an eye toward backward compatibility, as there is already an active user base.

C.7.1.6. *Computational Workflow*—Considering a photo-z workflow consuming data and resources directly from the DAC, it might use Butler to access the catalog data. It should be able to read individual columns from the Parquet files. Examples of common photo-z inputs are magnitudes (or fluxes) and respective measurement errors, but for machine learning methods the list of inputs can be fairly diverse (e.g. to include shape parameters, concentration, Sersic index, and others). For the official photo-z tables, the ones to be included in the object catalog for the data release, the query would just select columns with no selections at rows level. For other science-driven photo-z runs to be done by users, it is expected to have pre-processing/cleaning or filtering based on e.g., signal-to-noise, quality flags, region selections (e.g. cone search), etc.

The photo-z algorithms to be part of the data releases are to be defined after discussions at the Photo-z Validation Cooperative[22]. There is already a shortlist of algorithms to be tested available in dmtn-049, based on the letters of recommendations collected by the Photo-z Coordination Group. As it is now, DESC has a workflow that contains the estimation of photo-zs included as one of its steps via the *Redshift Assessment Infrastructure Layers* (RAIL) software package (see link in references). In this case, the output PDFs might be already set to be represented as qp objects. As RAIL is publicly developed code, other cases outside of DESC can use (and extend) RAIL (by adding additional photo-z algorithms of interest to the code base, such as othose being developed for the Brazilian IDAC), and therefore take advantage of each algorithm's native output formats. Alternatively, any other photo-z pipeline can also import qp at the end of the workflow and provide the parametrized PDFs as qp objects as well. Besides the main photo-z tables provided by DM, alternative photo-z tables will be provided by international contributors via the in-kind contribution program (e.g. LIneA in Brazil) as federated datasets, using the IDAC's infrastructure. Scalability tests are being done currently at LIneA using Parsl as workflow manager and Lephare and Color-Matched Nearest Neighbors (CMNN)[23] as examples of photo-z codes, using DESC's Cosmo DC2 mocks and DES data as precursor datasets. According to current results, it is estimated to take ~5 days to run CMNN for the whole LSST DR1, generating ~3 TB of outputs, without any post-processing (formatting/compressing) on the results. Although this estimate already fulfills the minimum requirements regarding computing speed (the order of milliseconds per object, as mentioned in https://dmtn-049.lsst.io/), even before planned optimization efforts begin, that estimate of computational time and storage footprint points to the necessity of adopting some strategy to transfer data in chunks in advance as soon as it is available, so it can be processed and made available as federated datasets, in the required output format, respecting the data releases timeline. As some

[22] https://community.lsst.org/t/rubin-commissioning-and-the-photo-z-validation-cooperative/6310
[23] https://github.com/dirac-institute/CMNN_Photoz_Estimator

subset of users will likely wish to run “custom" photo-z analyses on specific data subsets, it is worth considering that we should provide an “easy to use" photo-z pipeline to the community, with different algorithms available, providing formatted outputs, by default in the same formats as those official data products provided by DM with which people will already be familiar. In this case the photo-z code would be run by users on smaller subsets to address particular science cases (where for these cases the storage of results would either be local on the users’ own workspaces, or potentially on a shared resource). The easiest way to accomplish this would be to embed RAIL in the RSP and any IDAC infrastructures under consideration, however these options require development to distribute parallel jobs in different environments.

Parallelism, computation, storage and visualization: Many photo-z estimation codes are constructed in such a way that the analysis can be considered as embarrassingly parallel and also there are no constraints in spatial location of objects in the catalog. The partitioning can be optimized based on the machines’ memory availability.

There are no universal memory-per-core requirements, as these will be sensitive to choice of photo-z code and configuration adopted. Regarding visualization, qp includes some visualization tools but a more flexible and platform-independent solution would be desirable.

C.7.1.7. *References for Further Reading* —The qp software package for PDF storage can be found at https://github.com/LSSTDESC/qp. For 1D PDFs, Malz et al. (2018) is a good resource[24]. The LSSTDESC RAIL software package can be found at https://github.com/LSSTDESC/RAIL. A Roadmap to Photometric Redshifts for the LSST Object Catalog is available at Graham et al. (2022).

[24] Note that uses an older version of qp with additional though inefficient features.

C.7.2. *Photo-z: uncertainty quantification*

Contributors: Rachel Mandelbaum, Alex Malz, Raphael Shirley, Ashley Villar

C.7.2.1. *Abstract* —Most science cases that require a photo-z need some way to conveniently and quickly characterize the uncertainty in the photo-z. A p(z) (whether it is a posterior, a likelihood or some other distribution) can be many bytes of information; therefore, summary statistics describing these distributions are essential. For many science cases, we would like to answer questions such as:

- Is there a single "peak" in the p(z), or is it multimodal?
- What is the "best-fit" redshift or the peak redshifts in case of multimodality?
- What is the spread (second and higher moments)?
- Is there any statistical property (those listed or another, such as percentiles) of the p(z) that correlates with galaxy type?

Existing photo-z codes do not have a standardized way of reporting these user-defined summary statistics. We note one solution, RAIL[25], which standardizes photo-z outputs such as p(z) and can compute various point-estimated quantities from the p(z) using qp[26] as a back-end. RAIL does not necessarily calculate the quantities specified above, but it allows for easy implementation of such calculations.

We also note that "uncertainty" is a somewhat ambiguous term; what precisely are we quantifying? In this technical case, we only consider quantification of uncertainty *intrinsic* to photo-z estimates, whether the cause is epistemic (e.g. nonrepresentative spectroscopic training data or incomplete template library) or aleatoric (e.g. insufficiently complex estimation algorithm). Although a subtle difference, this is distinct from the question of the *accuracy* of the redshift estimates compared to reality.

C.7.2.2. *Science Cases Needing this Tool* —We enumerate a subset of science cases requiring photo-z uncertainty quantification:

- Extragalactic variable classification (AGN): Section B.4.3
- Extragalactic transient classification: Section B.3.1
- Cosmology (specifically, identification and characterization of tomographic samples might utilize uncertainty estimates): Section B.8
- Cosmology: cluster finders might use individual galaxy photo-z uncertainties when estimating cluster redshift and distinguishing cluster members from non-members sharing the line of sight: Section B.8
- Many static galaxy cases. For example, measuring the luminosity function of galaxies (especially as a function of redshift).: Section B.2.3

C.7.2.3. *Requirements for the software* —

- Timing:

[25] https://github.com/LSSTDESC/RAIL
[26] https://github.com/LSSTDESC/qp

  - Photo-zs must be produced for all of the galaxies (~5 billion) at least as often as every data release.
  - A nightly (re)calculation using multiwavelength data is needed for the alert broker-filtered extragalactic alerts in a given night.
  - Estimates must be updated as external information becomes available (e.g., as the spectroscopy training set expands, or with the addition of new SEDs to template libraries).
- Due to the number of galaxies, the software must be parallelizable.
- The memory requirements are (relatively) low. Lower-dimensional summary statistics must be much smaller than the p(z) itself (whose possible parameterizations are discussed in Section C.7.1), and there is no significant temporary/intermediate storage need.
- If storage parameterizations differ from usage parameterizations, software functionality for conversions must be provided.

Finally, we note that we can build these p(z) reduction tools today, or extend RAIL/qp to make them, and test on precursor datasets.

C.7.2.4. *Running on LSST and other Datasets* —We require the following:

- As stated above, we expect to be provided with p(z) information and summary statistics for all galaxies in the LSST object catalogs as well as p(z) for nearby galaxies in the nightly alert stream[27].
- In the first year of data (when the “official" p(z) will not be released), one could rely on a forced photometry measurement
- Ancillary data is required, in the form of spectroscopic training sets and SED template libraries; multiple versions thereof must be available for characterization of photo-z estimators.
- Before LSST, HSC or DES are helpful validation datasets.

C.7.2.5. *Existing Tools* —

- Most photo-z codes provide some characterization of uncertainty, however standardization of outputs is still missing.
- qp supports many formulations of uncertainty (e.g. moments) and can calculate point estimates from p(z); it could be extended to the specific ones requested here.
- Through the ELAsTiCC data challenge, brokers are being provided with p(z) quantiles in the alerts for the purpose of classification, along with a tool for converting the quantiles into other parameterizations (e.g. evaluations on a grid).

C.7.2.6. *Computational Workflow* —

[27] We note the p(z) information is considered proprietary and thus will not be in the public alert packet. Only the IDs for nearby Objects from the most recent DR will be included in the alert packet. Data-rights holders can use those to IDs query LSST databases for full p(z) information (via brokers or directly).

- Uncertainty quantification will involve accessing the photo-z PDFs for each individual extragalactic object in the object catalog, and doing some operations to the photo-z PDFs.
- This can be done in an embarrassingly parallel fashion.
- Results can be stored in a table. Visualization would typically involve exploring their correlation with object properties (magnitudes, colors, etc.)

C.7.2.7. *References for Further Reading* —

- Tanaka et al. (2018)
- Malz et al. (2018)
- RAIL for estimating p(z) and stress-testing estimators: https://github.com/LSSTDESC/RAIL
- qp for manipulating 1D PDFs: https://github.com/LSSTDESC/qp

C.7.3. *Photometric redshifts: Science driven metrics*

Contributors: Alex Malz (aimalz@nyu.edu), Colin Burke (colinjb2@illinois.edu), Raphael Shirley (r.a.b.shirley@soton.ac.uk), Andresa Campos (andresar@andrew.cmu.edu), Christa Gall (christa.gall@nbi.ku.dk)

C.7.3.1. *Abstract* —We propose a photometric redshift metric infrastructure targeted towards specific science cases that move beyond point estimate-based metrics. Derived population statistics (e.g. Stellar Mass Function (SMF) or luminosity function in a given redshift bin) should ideally have a performance metric associated with the standard LSST photo-z data products, including p(z). We propose each science collaboration with an interest in photo-z submit metrics which can be applied to the p(z) and to point estimate outputs of general photo-z estimators. This will permit public data challenges in addition to the development of new algorithms involving additional band measurements, imaging, positions, and other ancillary data.

- We want to move beyond simple, presumed Gaussian, errors for quantifying photo-z point estimates against spec-z measurements; instead, we want to apply metrics targeted to specific science cases (e.g. failure fractions for specific populations such as AGN, catastrophic outlier fractions as a function of magnitude and color, and galaxy population statistics for large samples of photo-z) that are sensitive to the quality of the photo-z uncertainty characterization.
- A set of available metrics will demonstrate the applicability of the provided photo-z information to a given science use case and be available for development and comparison of algorithms and input data options (e.g. additional bands, imaging, positions).

C.7.3.2. *Science Cases Needing this Tool* —

- Broad cosmological and extragalactic use cases including time-domain, transient, AGN, and galaxy science.
- Single object studies, selection based on photo-z derived values, and population studies all require targeted metrics.
- Various cosmological probes that incorporate photo-z information in different analysis stages, like weak lensing and galaxy clustering.

C.7.3.3. *Requirements for the software* —The most important metric theme that came out of the discussion was a notion of information; the accuracy of estimated uncertainties is not equivalent to the precision of an estimate. “True uncertainties" derived from a forward model of synthetic data are essential for comparisons to estimated uncertainties that must guide algorithm development to achieve the goal of accurate uncertainty characterization; such true uncertainties are due to limitations of the data itself rather than the model and are thus dependent on the galaxy population of interest. Generating true uncertainties for mock data requires software to model the space of redshift and data (whether the data is just LSST photometry or also images, positions, and other sources of photometry).

Such a model is only useful for metric evaluation if it includes “realistic complexity" to which the metric’s corresponding science case is sensitive, including selection effects and imperfect prior information, such as training sets and SED template libraries. The modeling should thus be flexible enough to include emission lines for AGN and other specialized galaxy populations to ensure at least one photo-z estimator performs well on such galaxies. Rather than being strictly based on extant data, the model should be extensible potential systematic differences of the types we should expect to discover with LSST (such as galaxy subpopulations in thus far unpopulated regions of color space).

For metrics that are only meaningful for specific subpopulations (such as AGN and transient classes), the software should be connected to subsampling or cross-matching software. That same connection would also enable the integration of data from outside sources, such as Roman or Euclid photometry, in running an estimator on subsets of galaxies. The software should then be able to save additional versions of the photo-z PDFs for those galaxies; not all galaxies in the catalog will have the same number of estimated photo-z data products, a complication the catalog’s format must accommodate.

Metrics needed:

- For transient and variable source classes (e.g. AGN), quality of match to external catalog of time-domain object identification and classification
- For each subpopulation of interest, flag for confidence in method based on estimation algorithm’s general performance on that subpopulation (e.g. AGN will require trustworthy uncertainty but only evaluated on Y1 AGN, by strength relative to host).
- Ensemble metrics relative to a higher-fidelity (spectroscopic) sample.
- Quantification of information content between multiple photometric redshift estimators (or same estimator with different priors, or additional bands, etc.)
- Connection to external pipeline for science metric, e.g. cosmology 3x2pt

However, ideally it should be possible to accommodate other key science-driven metrics.

C.7.3.4. *Running on LSST and other Datasets —*

- Spectroscopic training sets and/or SED template libraries will be necessary as prior information for photo-z estimators
- Initially, photo-z estimation will be performed using only the Rubin photometric catalogs, though it is likely that ancillary data (see below) will become more important as the survey progresses.
- Ancillary data where available, including imaging and/or more bands from other instrument(s) and/or positions to assess how additional bands contribute to photo-z accuracy.
- Deblended galaxy model parameters (e.g., Sérsic parameters) for every galaxy (at scale for photo-z) rather than image cutouts

C.7.3.5. *Existing Tools —*RAIL includes an extensible emulation suite for obtaining true posteriors conditioned on photometry to compare to estimates. Further development would be

necessary to interface with cross-matching/subsampling, to extend emulation to additional forms of data (imaging, more bands, positions, etc.), and to ensure realistic complexity for rare subpopulations. While the first of these may be straightforward, the latter two would require significant software development.

RAIL also includes an extensible framework for metrics, to which science case-specific metrics would have to be added.

C.7.3.6. *Computational Workflow* —

- Forward modeling of data (photometry and ancillary) with corresponding true PDFs to compare to estimates
- Training (or otherwise informing, for template-fitters) of estimators
- Estimation of photo-z PDFs on synthetic forward-modeled data
- Evaluation and comparison of metrics of point estimates and PDFs, including on specific subpopulations individually
- Evaluation and comparison of metrics computed across sky using random (or binned, such as by AGN strength relative to host) samples of objects

C.7.3.7. *References for Further Reading* —

- Using image information / CNN: Pasquet et al. (2019)
- AGN photo-z: Brescia et al. (2019)
- Galaxies photoz use case: https://community.lsst.org/t/lor-the-galaxies-science-collaboration-photo-z-use-case/5887
- AGN roadmap: https://agn.science.lsst.org/sites/default/files/LSST_AGN_SC_Roadmap_v1p0.pdf
- RAIL: https://github.com/LSSTDESC/RAIL

C.8. *Other technical use cases*

A few independently developed use cases were also submitted which did not fit in the broad areas outlined in Section C.1.

C.8.1. *Joint Calibration of precursor surveys for longer-baseline Light Curve Generation*

Contributors: Weixiang Yu (editor; wy73@drexel.edu), Colin J. Burke (colinjb2@illinois.edu), K.E. Saavik Ford (sford@amnh.org)

C.8.1.1. *Abstract* —Rubin LSST is unprecedented in its unique combination of depth, spatial coverage, and time-domain capability. However, many science cases could not take full advantage of LSST's time-domain capability until later into the survey, because some classes of objects (e.g., AGN, Mira, etc.) exhibit intrinsic long-term variability and a years-long (even decade-long) baseline is needed to classify and characterize them. Thus, we propose to jump start LSST variable sciences through joint calibration of current/past time-domain surveys with LSST and producing/serving forced-photometry light curves for variable LSST sources. Those re-calibrated archival light curves will not only enable early variable science for LSST sources, but also ensure a timely classification of variable sources into different classes (e.g., AGN vs. variable stars). A more complete & less contaminated AGN catalog is critical for MMA and CSQ searches.

C.8.1.2. *Science Cases Needing this Tool* —

- Early classification of variable sources in LSST
- Early variable science with LSST
- Statistical studies of certain classes of objects where a long baseline is needed to produce a complete and pure sample.
- Long-term structure function determination / AGN (non)-stationarity (Stripe 82)
- Rapid identification of a cleaner/more complete sample of AGN for searching for MMA counterparts, changing-state' quasars (CSQs) means identification of AGN-driven fraction of Laser Interferometer Gravitational-Wave Observatory (LIGO) sources, disk structure and astrophysics of turn on/off of accretion disks.

C.8.1.3. *Requirements for the software* —

- Careful cross-calibration of data from multiple surveys conducted with different hardware and at various sites.
- Generate/store forced-photometry light curves from re-calibrated archival data at the locations of variable LSST sources
- Extract time-series features from those light curves and store them in a database
- Preferably run this analysis a few times throughout the 10-year LSST survey.
- Visualization tools for serving those light curves to the end users through the RSP will be useful.

C.8.1.4. *Running on LSST and other Datasets —*

- The proposed work will only utilize data release catalogs. Ideally, the first run should be carried out once the variable nature of LSST sources can be reliably determined (primarily for saving computing resources).
- We can test the pipeline now on precursor data: HSC + BlackGEM/DECam/ZTF/PTF/Pan-STARRS/SDSS.
- Because commissioning fields will be chosen for overlap with precursor surveys, this procedure will be possible on Data Preview 2 (the LSSTCam commissioning data release)
- The data products resulting from running such a pipeline on precursor data sets will enable many new investigations that can benefit from a longer light curve baseline (certainly for AGN variability science).

C.8.1.5. *Existing Tools —*

- Light curves from different surveys can be merged by mean/median. However, without careful cross-calibration and color-term correction, the resulting light curves are not reliable, especially for sources exhibiting complex variability properties (non-stationarity on short time scales).

C.8.1.6. *Computational Workflow —*

- Cross-calibrate archival data from precursor surveys with LSST.
- Run forced-photometry on re-calibrated archival single-epoch images at the locations of variable LSST sources, color-correct the photometry, and store those light curves on disk for later retrieval.
- Extract time-series features from those light curves and store them in a database.

C.8.1.7. *References for Further Reading* —Longer baseline justification: Kozłowski (2017, 2021)
Photometric Calibration of PTF: Ofek et al. (2012)
Photometric Calibration of ZTF: https://irsa.ipac.caltech.edu/data/ZTF/docs/ztf_extended_cautionary_notes.pdf
Photometric Calibration of Pan-STARRS: Schlafly et al. (2012)
Photometric Calibration of DES and Rubin LSST: Burke et al. (2018)

C.8.2. *Multi-Stream Transient Detection and Characterisation*

Contributors: Tyler Pritchard (TylerAPritchard@gmail.com), Alex Gagliano, Samuel Wyatt, Igor Andreoni, Tomas Ahumada, Catarina Alves, Christa Gall, Suvi Gezari, Jing Lu, Fabio Ragosta, Clare Saunders, Adam Scott, Ashley Villar, Sam Wyatt, Ann Zabludoff

C.8.2.1. *Abstract*—LSST will discover millions of variable and transient events per night; the characterization of most of these objects will occur on longer timescales. With a likely single filter detection and inter-night revisit timescales of ~3–5 days, it will take multiple nights to characterize any individual transient with only LSST data. Individual groups will have their own follow-up resources; the optimal allocation of these facilities remains an open problem. One tool that could enable more prompt characterization, and help groups efficiently target their follow-up, is the correlation of multiple streams of public time-domain information on a single platform that would allow humans and/or algorithms to select targets and assign resources.

For example, one vision of this could be a broker-like interface with shared data from LSST, ZTF, ASAS-SN, and Gaia. A more comprehensive version could include publicly reported spectra that are correlated with known alerts (such as those provided by the TNS). This would result in products such as early estimated colors, if for example an LSST detection in a single filter is supplemented by detections in ZTF, ASAS-SN, or Gaia. Or, rather than different filters, multiple observations in a similar filter between LSST and another survey would result in an estimated magnitude difference between the two observations, and therefore brightness rate-of-change, before LSST or any single survey would be able to provide it. This could be useful across a range of science cases including the detection of very young or fast evolving explosive transients, X-ray binaries transitioning into an outburst, or changing look AGN.

Current limitations are primarily driven by the data volume and velocity of any single stream, as well as interface issues with a number of (theoretically at least) public data. One current solution that members of the transient community currently use is to create their own jupyter notebooks that connect to multiple data servers. This solution faces several technical issues including: queue-rate limits, difficulty of creating queries across different formats that are sufficiently cross-matched, local storage and compute, and a significant amount of custom pre-processing. This is difficult to scale today, and will become increasingly difficult without a unified platform as LSST comes online. Similar work is being done today in curated transient surveys, with less efficient tools or data releases, such as the Young Supernovae Experiment (the special ZTF follow-up of the TESS survey region with Pan-STARRS),the Global Supernova Project (an LCO follow-up of ZTF detections with an open/free-to-join community), Gaia alerts and third party spectroscopy made available on, e.g, the TNS. This need will only grow in the future as more overlapping surveys are planned -including the search for transients in DESI and the Roman Observatory, future proposed rapid-transient DECam surveys in the era of the Rubin Observatory LSST.

This shares overlap with the multiple-instrument algorithms, but is focused on derived-products such as light curves and photo-z's (and potentially additional metadata such as classification outputs).

C.8.2.2. *Science Cases Needing this Tool*—Example science cases that would benefit from this tool include:

- Anomaly Detection & Follow-up – things can be anomalous along multiple axis, and follow-up is often required as early as possible
- X-ray Binary Outbursts – the shape and timing of the outburst provides information about the mass and size of the binary system.
- Fast Evolving Transients & Early Supernovae – LSST will provide a deluge of newly discovered transients. Identifying the select few that merit follow-up observations will prove challenging, especially as the information from LSST will be limited. With a latency of several days between repeated observations in the same filter, it will be difficult to acquire follow-up early in the evolution of newly discovered LSST transients.
- Changing-look AGN – these systems evolve between observational AGN types on rapid timescales.
- Quantifying sample purity, contamination, completeness for e.g., SN Ia cosmology.

We also note that this tool would be useful for "archival" studies, particularly the development of multiple-instrument ML algorithms.

C.8.2.3. *Requirements for the software*—The requirements for such a tool share many similarities with that of an LSST broker, but emphasize the combination of multiple streams and the need for a rapid response.

- The ability to combine multiple kafka or LSST-alerts streams from publicly available sources (including multi-wavelength or multi-messenger surveys). This could be across the entire survey footprint for multiple streams, or constrained to a well-defined overlap between two streams for a more narrow use case.
- Ideally, the ability for *relatively small* surveys or observing programs to easily contribute data, potentially for something like known events, post-detection.
- The ability to search for and filter off of cross-matched alert sources - either from each individual stream or in the combined data-set.
- The ability to cross-match or query other catalog or annotated information to enable prompt decision making.
- Rapid publication of the cross-matched streams for greatest impact.
- The biggest challenge is primarily the volume and velocity of data – while the LSST alert stream will dominate the volume of the data, other sources will expand the scope of what is stored significantly. If this can be done in a timely manner, it will help inform the community develop plans for follow-up.

- While more challenging, it would be good for the service to additionally annotate data by providing host galaxy cross-matching from catalogs (e.g., Glade or GWGC3) or spectra that get publicly reported.

C.8.2.4. *Running on LSST and other Datasets* —

- This would be focused on the LSST Alert-stream and could begin immediately upon survey start.
- The LSST Alert stream data can then be combined with both extant and future public time-domain data (e.g., ZTF, ASAS-SN, Gaia, future DECam Surveys, Roman, other public data).
- Scaled down test surveys are productive and being conducted now (e.g., the Young Supernovae Experiment), and in a more limited sense with other surveys incorporating ZTF data into their transient detection/characterisation methods (e.g., DLT40, The Asteroid Terrestrial-impact Last (ATLAS)).

C.8.2.5. *Existing Tools* —The tools needed largerly exist as the functionality overlaps with the needs of an LSST Alert broker, including:

- Big Data storage (e.g., postgresql/NoSQL/cloud buckets)
- Methods to rapidly and efficiently stream data (e.g., Kafka, pub-sub systems, and shared cloud buckets)
- HyperText Transfer Protocol (HTTP) API access with potential web visualization interface

What's missing is the expansion and publication of this to multiple combined data streams, such as:

- Methods to efficiently cross-match positional data between different alert streams ZTF, ASAS-SN, and others (e.g., postgris World Wide Web Consortium (W3C) spatial tools & ToPCAT), AXS (Astronomy Extensions for Spark).
- For events with large positional uncertainties and time-critical follow-up, methods to publish observation details for community coordination such as treasuremap.space.
- Methods for users to filter off of stream data including personal jupyter notebooks doing small scale processing on curated streams, Target Opportunity Managers (TOM), and broker filter interfaces.

C.8.2.6. *Computational Workflow* —This will vary depending upon the chosen methodology, number of surveys or data streams, scientific interest, and wavelength(s), but one workflow for a minimal implementation for a specific science case could be:

- Identify a target list. For extra-galactic transients, for example, this could be the TNS, for variable stars the AAVSO Variable Star Index, for AGN a joint Gaia-WISE curated list, or a LIGO or IceCube alert stream.
- Query photometry/alert servers for data upon the publication of a new event (e.g., TNS, LSST/ZTF brokers, ASAS-SN Sky Patrol, Gaia.

- In the case of events with open data but no published photometry stream (e.g., Swift, Fermi), the stream could be annotated to indicate where observations exist even if detections/limits are not available.
- (optional) A feedback mechanism for vetted groups without a dedicated photometry stream to provide data on events. This could be potentially supplementary (e.g. spectra, a redshift, classifications or periods) or additional photometry.
- A re-publication (on-line or via api) of the Alert stream with the additional observation or annotated list of observations).

A larger, more science agnostic case could be:

- Take a curated Kafka stream from an LSST broker focusing on non-Solar System Object (SSO) transient events and a second stream from a second survey (potentially in a defined sub-survey region, e.g., DDF or overlap with a DECam Survey)
- Spatially cross-match the two streams
- (Optional) Add additional streams if available
- (Optional) Add annotated information including spectra, epochs of observations by other telescopes/wavelengths (potentially archival/long baseline), classifications, periods, and other information.
- Publish a joint stream with the combined data to an end user, either through a broker, TOM, or pub-sub channel.

This could be scaled up depending on overlapping observation width, survey period, and number of streams.

C.8.2.7. *References for Further Reading*—Brokers, Tools, Sources mentioned above: Alerce Antares Fink, Lasair, Transient Name Server, AAVSO Variable Star Index, Gaia Alerts, Young Supernova Experiment, ASAS-SN SkyPatrol, Global Supernovae Project Supernova Exchange, Treasure Map, TOM Toolkit, Astronomy Extensions for Spark

C.8.3. *Interactive Data Visualization at scale*

Contributors: Leanne Guy (leanne.guy@lsst.org), Ilija Medan (imedan1@gsu.edu), Jing Lu (jl16x@my.fsu.edu), Tomislav Jurkic (tjurkic@phy.uniri.hr), Juan Luna (jmluna@iafe.uba.ar), Tomas Ahumada (tahumada@astro.umd.edu), Rosaria (Sara) Bonito (rosaria.bonito@inaf.it), Sabina Ustamujic (sabina.ustamujic@inaf.it), Markus Hundertmark (markus.hundertmark@uni-heidelberg.de), Yiannis Tsapras (ytsapras@ari.uni-heidelberg.de), Matthew Graham, Neven Caplar, Ashish Mahabal, Mark Popinchalk, Viviana Acquaviva, Catarina S. Alves, Garrett Levine

C.8.3.1. *Abstract*—At the end of the 10-year LSST, the size of the final Object Catalog is expected to be approximately 15PB in size and contain approximately 40 billion Objects. Developing a framework for automated interactive visualisation of the LSST dataset is key to discovery. The classic technique of subsetting the data down to a manageable size that will fit into memory, by defining cuts on various object attributes and then visualising the resultant data, limits discovery potential by reducing the dataset based on assumptions about the data. Creating visualisations based on the full dataset will allow scientists to explore the full LSST discovery space interactively. All LSST science domains can benefit from visualisation that enables interactive exploration, subsetting, drilldown, and brushing and linking between plots. In 2016, the Gaia Data Release 1 (DR1) density map of over 1 billion sources in the Milky Way was named "One of the five coolest things on Earth this week" by General Electric. Many industry-standard tools exist already that we can take advantage of for creating powerful visualisations of peta-scale datasets, e.g Holoviz (https://holoviz.org/). LSST can benefit enormously by building upon these frameworks and tools.

C.8.3.2. *Science Cases Needing this Tool*—All science domains and use cases can benefit from powerful interactive visualisations. Some notable examples include:

- Visualizing the near real-time light curves generated from custom cutouts of images. This will be useful for a variety of extended objects, e.g. galaxies, solar system objects, young stellar objects, as well as transient events such as gravitational microlensing events, SN, and GRBs in particular in crowded fields.
- Studying accretion; rapid variability on timescales from minutes to hours or days is a hallmark of accreting systems, from compact binaries to AGNs and young stellar objects (YSOs). The time scale and amplitude of the variability provide information about flares, rotation, and the accretion process including its rate and power. Being able to obtain quick on-the-fly measurements of the variability of a source, such as light curve variance, would enable rapid identification. The nature of the variable sources could then be further narrowed-down with a (linked) visualisation of their location on color-color or color magnitude diagrams.
- Anomaly detection for young/unknown sources.
- Density maps.
- Investigating the attributes of points in linked plots in order to drill down to images, which could be retrieved dynamically via the cutout server, in order to perform further

analyses (e.g. finding and relating host galaxies to transients and to visualize them in the LSST images).

- Providing multi-level linked selection of aggregated data to look for correlations between parameters, e.g computed scientific parameters or latent parameters following dimension reduction. An example of the progression of levels could be "Sample subsets -> correlation heat map of parameters -> individual x-y plot", where these can be combined with drill down to zoom in on the individual target.
- On-the-fly calculation of basic statistics or other quantities from selected areas on plots, e.g computing the mean and median of a selected a region on spatial distribution plot.
- 3D rendering of select Objects, such as YSOs, whose light curves exhibit varying shapes, e.g. a dip due to warp disks or modulation due to rotation or variability associated with accretion/ejection processes, using external tools such as paraview for data analysis or the sketchfab platform to share models interactively. The 3D rendering will allow us to explore different line-of-sights and to understand how geometric effects change the observed light curves. The TVSSC is starting a program on 3D visualization.
- Mapping a World Coordinate System (WCS) onto a plot.
- Producing interactive dashboards that enable exploration of the parameter space of a given class of object. For example, tweaking SN Ia parameters to fit light-curves.
- Providing an interface to interactively perform period finding and phase folding for light curves. For example, a slider bar for changing the period that would adaptively phase fold the light curve to that period.
- Provide easy to use Observing Program Management Systems to assist with managing and tracking active observing programs. The TVSSC has solicited an in-kind contribution for this purpose.

C.8.3.3. *Requirements for the software*—There are many excellent open-source tools on the market for producing stunning visualizations. We should not think about writing a visualization tool, rather we should adopt an existing tool or a framework and write software to provide an interface between these industry standard tools and the LSST data products. Producing visualizations of large datasets may require additional CPU or GPU or the use of a tool like Dask to compute the quantities to visualize. This approach provides a highly optimized rendering pipeline that makes it practical to work with extremely large datasets even on standard hardware, while exploiting distributed and GPU systems when available.

There is no reason why we cannot start to build these tools today. The biggest impediment is likely to be that people are not used to this manner of working and will require training. This requires something of a change of culture.

Requirements:

- Rapidly create visualizations of different classifications of objects, e.g stars vs. galaxies.

- Link plots that can present different representations of the same data.
- Brushing and linking between different visualizations of the same data points.
- Speed for rapid visualization.
- Good annotation of data across multiple surveys to allow for plotting/visualisation of data on an object from all of these surveys.
- Drilldown capabilities (images/spectra from other surveys by cross-matching)
- Integration with external tools (when they can be installed on the science platform), e.g creating 3D models.
- Able to integrate new software that can calculate needed derived quantities (e.g. statistics on data, representation in healpix) relevant to science use cases
- Adding domain specific knowledge (i.e. overlaying WCS) onto industry specific software
- Ability to transform data from LSST tables/parquet files into a data format is needed by the visualization tool.
- Ability to automatically and accurately render billions of objects (e.g Datashader) rapidly and flexibly including annotations and interactive capabilities.
- Need to produce static images from final plots to include in papers
- Any third-party software that we use should be open-source.
- Rendering billion point datasets could require large processing power or GPU
- Traditionally people will store a pdf/png of the final plot. This means that to recreate the plot (say in order to change colours or point size), all the computations need to be redone to recreate the plot. This can be time consuming. It may be preferable to store the aggregated data resulting from the computations that go into making the plot. This way the plot can easily be recreated to change plot attributes.
- Ability to align lightcurve data of a single target from different sources on a visualization.
- Ability to interactively mask datapoints
- Include options/capabilities for those that are e.g. visually impaired.
  - Developing 3D printed kits to include visually impaired scientists and to be used for scientific dissemination.
  - Data sonification

C.8.3.4. *Running on LSST and other Datasets*—Scientists will need to visualize the following data products:

- Object catalog and ForcedSource catalogs to visualize time series and associate object attributes
- Full visit images or cutouts of the LSST images obtained via the cutout server
- LSST data combined with data from various external catalogs in a single visualization. Note that data access might require the use of queries over TAP and of the services of a cross-match service.

- Precursor and simulated datasets such as DES DRX, HSC PDRX, DESC DC2. can serve to develop and validate visualization tools in preparation for LSST data and possibly lead to publications if advanced visualization of these datasets leads to new discovery. The RSP provides an ideal environment to run on and develop visualizations.
- Objects with selections based on HEALPix maps

C.8.3.5. *Existing Tools* —Existing tools for processing/visualising data include:

- lightkurve (inherits partially from Astropy's TImeSeries)
- The Holoviz suite of tools including Bokeh, Holoviews and datashader and panel.
- Altair to link plots
- Treasuremap
- Plotly for dashboards
- Paraview[28] data analysis and visualization application. It can be run on supercomputers to analyse extremely large datasets also with Python[29]

Many of these tools are used in industry to visualize large datasets and are expected to work at LSST scale. Holoviz has been demonstrated on the DESC Data Challenge 2 (DESC) (DC2) dataset via Rubin DP0. What is needed is to provide the interface layer to the LSST data products.

C.8.3.6. *Computational Workflow* —

- Input data to a visualization framework would be either a) directly from parquet files provided by the project, b) from results of ADQL queries of LSST databases, or c) images retrieved either via the Butler or cutout server. Additionally, it would be useful to work with community alert brokers to visualize data in alert packets.
- Steps include: Data preparation, including extracting data/parameters from processed images, parametric fitting, classification, dimension reduction, joining with results of cross-matching,
- We would expect to be able to use the project provided batch and reprocessing tools for any reprocessing.
- Creating density maps from the full catalog data will require distributed processing frameworks such as Apache spark or Dask to aggregate, bin or compute other derived quantities for visualization.
- Outputs could be stored in user databases to avoid recomputation if computation cost is high.

C.8.3.7. *References for Further Reading* —

- Astropy (Astropy Collaboration et al. 2013, 2018)
- Holoviz[30]

[28] https://www.paraview.org
[29] https://www.paraview.org/python
[30] https://discourse.holoviz.org/

- Firefly[31]
- Matplotlib (Hunter 2007)
- Lightkurve (Lightkurve Collaboration et al. 2018)
- Altair[32]
- TreasureMap (Wyatt et al. 2020)
- yt (Turk et al. 2011)
- GlueViz (Beaumont et al. 2015; Robitaille et al. 2017)
- Houdini[33] (e.g. Naiman et al. 2017)
- TOPCAT (Taylor 2005)

[31] https://github.com/Caltech-IPAC/firefly
[32] https://altair-viz.github.io/index.html
[33] http://www.ytini.com/

D. SCENARIOS USED FOR THE INCLUSIVE COLLABORATION BREAKOUTS

The following four scenarios were used in the breakout sessions focused on inclusive collaboration. Participants were deliberately split into diverse groups. Each group did a round table of introductions, then read two of the scenarios and discussed them.

The goal was stated as "People come to these discussions with their own experiences and backgrounds. There may be some who have experiences very similar to the scenarios. Today's goal is to focus on the scenarios and to identify productive strategies based on the prompts that we all could potentially use to foster positive collaborations and avoid repeating prior mistakes."

D.1. *Scenario 1 – Institutional Pressures*

Zahra is part of a large research team at an Doctoral Universities – Very high research activity (R1) (research intensive) institution developing a Research Inclusion plan for their research proposal. The team approaches Carl, a teaching institution astronomer that Zahra knows from a past AAS meeting to join the team. Zahra thought Carl would be a good fit for the team because his dissertation research was on a similar topic as the topic they will study should their proposal get accepted. Zahra emails Carl asking if he is interested in joining their team. Carl realizes this could be a good opportunity, as he is expected to publish (albeit minimally relative to an R1 institution) to qualify for tenure. However, Carl is apprehensive about joining the team because he doesn't have much experience in large collaborations. He is also worried about the different institutional pressures they face as they work for different types of institutions.

Questions:

1. What are some different pressures that researchers from research-intensive and teaching institutions may face?
2. What are some steps that Zahra's collaborative team can take to make Carl's participation on the team valuable for him?
3. What conversations could Zahra and Carl have during these early stages to better understand different institutional contexts, collaboration expectations, and collaborator capacities?
4. What questions might Zahra or Carl ask each other to begin laying a foundation for a successful collaboration?
5. Have you ever participated in a collaboration that spans multiple institutions of different sizes or types? What worked well in those collaborations? What was challenging? If tensions arose based on different institution types, how were these tensions settled or resolved (if at all)?

D.2. *Scenario 2 – Allocation of Credit*

A collaboration centered at a few large research institutions approaches Sarah, an assistant professor of astronomy at a teaching institution, to ask if she is interested in collaborating on a proposal. The team explains that Sarah would be a good fit because of her expertise, and

her involvement in the collaboration would contribute to the institutional diversity of the project as well. Sarah's involvement would become a central pillar of a required research inclusion component to the project, which would make the proposal more competitive. Sarah thinks this could be a great opportunity because she needs to publish more articles before going up for tenure. She knows that research collaboration tends to be more highly cited, more visible, more innovative, and more likely to have a greater impact than sole authored research. Funders often view collaboration more favorably than sole-investigator research as well. This opportunity could really strengthen her tenure file. However, Sarah is also worried that it will be more difficult to get credit for her component of the work on such a large project. Sarah's hesitation is exacerbated when in preliminary discussions the team plans authorship on their first planned publication; a postdoc at the research-intensive university with similar expertise as Sarah is slated to be listed above her in authorship despite both of them being equally important to the development of the manuscript.

Questions:

1. How and when do you typically navigate authorship on research publications? What strategies do you have for managing tensions that may arise around authorship?
2. Why might the team in the example above assume the postdoc would be higher in authorship order than Sarah even if their contributions are roughly equal?
3. What are some ways the team could ensure the collaboration will be meaningful for Sarah? What may be meaningful for Sarah's type of institution? What types of conversations could the team have to illuminate what is meaningful for each collaborator, including Sarah, given their institutional contexts?
4. What are some ways the team could ensure collaborators are properly credited for their contributions?
5. What are some useful team publication conversations to have during this "inviting collaborator" stage? What issues might they address?
6. Have you ever had a negative research experience that arose due to the allocation of credit on a collaborative paper? If so, what lessons did you learn from that experience?

D.3. *Scenario 3 – Inclusive Team Environment*

Wei, an astrophysicist, is leading the development of a research proposal that (if accepted) would span research and teaching institutions, and consist of team members from different career stages (e.g. associate professors, assistant professors, postdocs, graduate students). The proposal requires a Research Inclusion plan alongside the Science and Data Management plans. In drafting the Research Inclusion plan, Wei is prompted to describe how the team can foster an inclusive environment in which all team members feel respected and valued for their unique contributions to the project.

Questions:

1. What are some examples of team dynamics that may foster an environment that is isolating to, or excludes, some members of the team?

2. What are some strategies that Wei and other team members could employ to foster an inclusive environment in which researchers of different ranks feel valued on the project? What are effective ways to implement these strategies?
3. What are some strategies that Wei and other team members could employ to foster an inclusive environment in which researchers from different types of institutions feel valued on the project? What are effective ways to implement these strategies?
4. Have you ever been on a team (or part of a department) in which the environment was not inclusive of everyone? How were people excluded or made to feel like they were not fully integrated into that space? What were the impacts of the exclusionary environment? In what ways were these negative dynamics addressed (if at all)?

D.4. *Scenario 4 – Student Contributions to Open-Source Software*

An astronomer working at a teaching-intensive institution, named Ahmed, is collaborating with a team of astronomers from large, research-intensive institutions to build some open-source software for data reduction. As part of this work, Ahmed wants to include some undergraduates from his institution on the project, as the project would provide these undergraduates with research and coding experiences typically unavailable to them. Including undergraduates would also benefit Ahmed; the administrators at Ahmed's institution reward this type of student engagement in tenure decisions.

Questions:

1. What advice would you give those organizing the software development effort to provide a clear and welcoming path to involvement from new contributors such as Ahmed's students? For example, this might involve providing simple explanatory references, reducing use of jargon in the documentation and clearly defining it when it appears, or other measures.
2. How can Ahmed and the entire team ensure that students who invest significant short-term effort into developing open source software as part of this code base get credit for their contributions?
3. How does the community ascribe credit for widely used, open-source software? What author-order considerations should be taken into account for software development credit? How does software development credit differ from other authorship credit scenarios? What about open-source software development?
4. When does writing a piece of open-source software convey authorship? Are there processes we can put in place to help guide this decision? How does software infrastructure (e.g. running software at scale) fit into this picture?

E. **Glossary**

1D: One-dimensional. 70, 72, 213, 215

2MASS: Two-Micron All Sky Survey. 102

3D: Three-dimensional. 119, 229, 230

AAS: American Astronomical Society. 12, 233

ACT: Atacama Cosmology Telescope, for CMB observations. 167, 168

active asteroid: small Solar System bodies that have asteroid-like orbits but show comet-like visual characteristics. 137, 142

ADAM: Asteroid Discovery, Analysis, and Mapping. 132, 134

ADQL: Astronomical Data Query Language. 192

AGN: Active Galactic Nuclei. 4, 9, 40, 57, 64, 65, 68–70, 72, 73, 75–77, 79–83, 177, 181, 183, 186, 188, 192, 196, 206, 216, 219–222, 224, 225, 228, 239

AI: Artificial Intelligence. 105

ALeRCE: Automatic Learning for the Rapid Classification of Events. 61

Alert: A packet of information for each source detected with signal-to-noise ratio > 5 in a difference image by Alert Production, containing measurement and characterization parameters based on the past 12 months of LSST observations plus small cutouts of the single-visit, template, and difference images, distributed via the internet. 64, 67, 117, 119, 130, 133, 140, 226, 227

algorithm: A computational implementation of a calculation or some method of processing. 9, 46, 67, 70, 71, 77, 79, 92, 109, 132, 133, 138, 151, 158, 165, 174, 191, 209, 247

ALMA: Atacama Large Millimeter Array (ESO). 148, 238

AMPEL: Alert Management, Photometry, and Evaluation of Light curves. 65

ANTARES: Arizona-NOIRLab Temporal Analysis and Response to Events System. 55, 58, 115

API: Application Programming Interface. 203, 226

ARAS: Astronomical Ring for Access to Spectroscopy. 103

arcmin: arcminute minute of arc (unit of angle). 201

Arizona-NOIRLab Temporal Analysis and Response to Events System: ANTARES is a real-time astronomy system under development at NOIRLab. https://antares.noirlab.edu. 55, 236

ASAS-SN: All-Sky Automated Survey for Supernovae. 71, 224, 226

Asteroid Discovery, Analysis, and Mapping: a cloud-based astrodynamics platform in development by the Asteroid Institute, a program of the B612 Foundation. 132, 236

astrometry: In astronomy, the sub-discipline of astrometry concerns precision measurement of positions (at a reference epoch), and real and apparent motions of astrophysical objects. Real motion means 3-D motions of the object with respect to an inertial reference frame; apparent motions are an artifact of the motion of the Earth. Astrometry per se is sometimes confused with the act of determining a World Coordinate System (WCS), which is a functional characterization of the mapping from pixels in

an image or spectrum to world coordinate such as (RA, Dec) or wavelength. 34, 94, 96, 98, 99, 209

ATLAS: The Asteroid Terrestrial-impact Last. 226

Automatic Learning for the Rapid Classification of Events: The ALeRCE broker is a Chilean-led broker which is processing the alert stream from the ZTF and a Community Broker for the Vera C. Rubin Observatory and its LSST, as well as other large etendue survey telescopes. http://alerce.science/. 61, 236

Avro: is a row-oriented remote procedure call and data serialization framework developed within Apache's Hadoop project. 65

B: Byte (8 bit). 142

background: In an image, the background consists of contributions from the sky (e.g., clouds or scattered moonlight), and from the telescope and camera optics, which must be distinguished from the astrophysical background. The sky and instrumental backgrounds are characterized and removed by the LSST processing software using a low-order spatial function whose coefficients are recorded in the image metadata. 34, 45, 64, 139, 150

BEAMS: Bayesian Estimation Applied to Multiple Species (software for classification of light curves based on photometry). 160, 161

BH: Black Hole. 82

BHB: Blue Horizontal Branch. 89

BHNS: Black hole-neutron star. 62

BlackGEM: is a wide-field array of optical telescopes to be located at ESO's La Silla Observatory in Chile's Atacama desert.. 74, 108, 109, 123, 223

Blazhko: the phenomenon of amplitude or phase modulation. Associated with some RRL. 126

BNS: Binary Neutron Star. 62

BOSS: Baryon Oscillation Spectroscopic Survey. 168

Broker: Software which receives and redistributes Alerts, and may also perform processing such as filtering for certain characteristics, cross-matching with non-LSST catalogs, and/or light-curve classification, in order to identify and prioritize targets for follow-up and/or make scientific analyses.. 140

Butler: A middleware component for persisting and retrieving image datasets (raw or processed), calibration reference data, and catalogs. 214

CADC: Canadian Astronomy Data Centre. 140, 204

cadence: The sequence of pointings, visit exposures, and exposure durations performed over the course of a survey. 97, 113–117

calibration: The process of translating signals produced by a measuring instrument such as a telescope and camera into physical units such as flux, which are used for scientific analysis. Calibration removes most of the contributions to the signal from

environmental and instrumental factors, such that only the astronomical component remains. 88, 95, 96, 104, 157–159, 174

Camera: The LSST subsystem responsible for the 3.2-gigapixel LSST camera, which will take more than 800 panoramic images of the sky every night. SLAC leads a consortium of Department of Energy laboratories to design and build the camera sensors, optics, electronics, cryostat, filters and filter exchange mechanism, and camera control system. 96, 140, 239

camera: An imaging device mounted at a telescope focal plane, composed of optics, a shutter, a set of filters, and one or more sensors arranged in a focal plane array. 80

CARMA: Continuous time autoregressive moving average process, standard way to describe optical AGN variability. 73, 74, 77

CARMENES: Calar Alto high-Resolution search for M dwarfs with Exoearths with Near-infrared and optical Echelle Spectrographs. 94

CASA: Common Astronomy Software Applications (for ALMA). 148

CASLEO: Complejo Astronómico El Leoncito. 104

CCD: Charge-Coupled Device. 103, 127

CCL: Core Cosmology Library, https://github.com/LSSTDESC/CCL. 168, 174

CDF: Cumulative Distribution Function. 212

Center: An entity managed by AURA that is responsible for execution of a federally funded project. 132, 243

Charge-Coupled Device: a particular kind of solid-state sensor for detecting optical-band photons. It is composed of a 2-D array of pixels, and one or more read-out amplifiers. 103, 238

CHIME: Canadian Hydrogen Intensity Mapping Experiment. 62

Citizen Science: the collection and analysis of data relating to the natural world by members of the general public, typically as part of a collaborative project with professional scientists.. 141, 142

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).. 134, 226

CMASS: constant mass, a spectroscopic galaxy sample as part of the BOSS survey. 168

CMB: Cosmic Microwave Background. 3, 41, 156, 166–168, 236

CMB-S4: Cosmic Microwave Background Stage 4. 167

CMNN: Color-Matched Nearest Neighbors. 214

CNN: Convolutional Neural Network. 43, 44, 74, 127, 201, 208, 221

CNP: Conditional Neural Processes. 70–72

configuration: A task-specific set of configuration parameters, also called a 'config'. The config is read-only; once a task is constructed, the same configuration will be used to process all data. This makes the data processing more predictable: it does not depend

on the order in which items of data are processed. This is distinct from arguments or options, which are allowed to vary from one task invocation to the next. 215

Construction: The period during which LSST observatory facilities, components, hardware, and software are built, tested, integrated, and commissioned. Construction follows design and development and precedes operations. The LSST construction phase is funded through the NSF MREFC account. 164

CoRoT: Convection, Rotation et Transits planétaires. 117

CPU: Central Processing Unit. 115, 207, 229

CRTS: Catalina Real-Time Transient Survey. 35

CSM: Circum-Stellar Material. 59, 61

CSQ: Changing state quasar or AGN. 73, 75, 77, 222

CTTS: Classical T Tauri stars. 116

Cyber Infrastructure: Sometimes denoted CI, A term first used by the US NSF, and it typically is used to refer to information technology systems that provide particularly powerful and advanced capabilities.. 194

Data Management: The LSST Subsystem responsible for the Data Management System (DMS), which will capture, store, catalog, and serve the LSST dataset to the scientific community and public. The DM team is responsible for the DMS architecture, applications, middleware, infrastructure, algorithms, and Observatory Network Design. DM is a distributed team working at LSST and partner institutions, with the DM Subsystem Manager located at LSST headquarters in Tucson. 8, 234, 240

Data Release: The approximately annual reprocessing of all LSST data, and the installation of the resulting data products in the LSST Data Access Centers, which marks the start of the two-year proprietary period. 211, 240

DB: DataBase. 103, 127

DC2: Data Challenge 2 (DESC). 231

DDF: Deep Drilling Field. 53, 154, 227

DE: dark energy. 242

deblend: Deblending is the act of inferring the intensity profiles of two or more overlapping sources from a single footprint within an image. Source footprints may overlap in crowded fields, or where the astrophysical phenomena intrinsically overlap (e.g., a supernova embedded in an external galaxy), or by spatial co-incidence (e.g., an asteroid passing in front of a star). Deblending may make use of a priori information from images (e.g., deep CoAdds or visit images obtained in good seeing), from catalogs, or from models. A 'deblend' is commonly referred to in terms of 'parent' (total) and 'child' (component) objects. 34

DEC: Declination. 203

DECaLS: The Dark Energy Camera Legacy Survey. 140

DECam: Dark Energy Camera. 96, 140, 154, 223

DECAT: DECam Alliance for Transients. 154

DEEP: Deep Extragalactic Evolutionary Probe. 154

DELVE: DECam Local Volume Exploration Survey. 43, 89, 90

DES: Dark Energy Survey. 43, 56, 73, 74, 77, 83, 140, 156, 159, 168, 173–175, 212, 217, 223

DESC: Dark Energy Science Collaboration. 20, 157, 163, 171, 173, 174, 231, 239

DESI: Dark Energy Spectroscopic Instrument. 41, 43, 89, 168, 170–172, 174, 209

DHO: damped harmonic oscillator. 69

DIA: Difference Image Analysis. 139, 202, 205–207, 210

Difference Image Analysis: The detection and characterization of sources in the Difference Image that are above a configurable threshold, done as part of Alert Generation Pipeline. 139, 240

DM: Data Management. 8, 19, 20, 80, 127, 157

DP0: Data Preview 0. 210

DR: Data Release. 211

DR1: Data Release 1. 228

DR3: Data Release 3. 192

drill down: Move from a higher level aggregation of data to its inputs. For example, given data describing a tract, to drill down to constituent patches and then to objects. Also refers to the act of identifying an issue in a high-level summary of the data (e.g. an aberrant metric value) and interactively investigating its inputs to find the source of the problem. 228, 229

DRW: damped random walk. 69

EC2: Amazon Elastic Compute Cloud. 194

Education and Public Outreach: The LSST subsystem responsible for the cyberinfrastructure, user interfaces, and outreach programs necessary to connect educators, planetaria, citizen scientists, amateur astronomers, and the general public to the transformative LSST dataset. 209, 240

ELG: Emission-Line Galaxies. 171

EPO: Education and Public Outreach. 209

epoch: Sky coordinate reference frame, e.g., J2000. Alternatively refers to a single observation (usually photometric, can be multi-band) of a variable source. 201, 202, 209

ESA: European Space Agency. 137

ESO: European Southern Observatory. 148, 236

FBOT: Fast blue optical transient. 60

FBOTs: Fast blue optical transients. 59, 186

FELTs: Fast-Evolving Luminous Transients. 59

Filter: A filter in astronomy is an optical element used to restrict the passband of light reaching the focal plane, it transmits a selected range of wavelengths. Filters elements are often named after standard photometric passbands, such as those used in the SDSS survey: u, g, r, i, z. 64, 65, 92

Firefly: A framework of software components written by IPAC for building web-based user interfaces to astronomical archives, through which data may be searched and retrieved, and viewed as FITS images, catalogs, and/or plots. Firefly tools will be integrated into the Science Platform. 232

FITS: Flexible Image Transport System. 106, 109

Flexible Image Transport System: an international standard in astronomy for storing images, tables, and metadata in disk files. See the IAU FITS Standard for details. 106, 241

flux: Shorthand for radiative flux, it is a measure of the transport of radiant energy per unit area per unit time. In astronomy this is usually expressed in cgs units: erg/cm2/s. 57, 75, 104, 116, 118, 119, 157

ForcedSource: DRP table resulting from forced photometry. 230

FOV: field of view. 64, 208, 241

FoV: Field of View (also denoted FOV). 64

FWHM: Full Width at Half-Maximum. 137, 138

Gaia: a space observatory of the European Space Agency, launched in 2013 and expected to operate until 2025. The spacecraft is designed for astrometry: measuring the positions, distances and motions of stars with unprecedented precision. 88–90, 120, 121, 124, 125, 130, 179

GALAH: GALactic Archaeology with HERMES. 88, 96

GAMA: Galaxy And Mass Assembly (survey). 43, 170, 171

GLADE: Galaxy List for the Advanced Detector Era. 63

GPU: Graphics Processing Unit. 72, 196, 208, 230

GR: General Relativity. 84

GRB: Gamma-Ray Burst. 55, 62

GSE: Gaia Sausage-Enceladus. 87, 88

GW: Gravitational Wave. 40, 62

GZ: Galaxy Zoo. 48, 207

HB: Horizontal Branch. 87

HEALPix: Hierarchical Equal-Area iso-Latitude Pixelisation. 90, 231

HELP: Herschel Extragalactic Legacy Project. 53

HERMES: a high-resolution fibre-fed spectrograph for the 1.2m Mercator telescope. 88, 241

HITS: High Cadence Transient Survey. 154

HSC: Hyper Suprime-Cam. 43, 53, 74, 140, 156, 158, 159, 168, 170, 171, 174, 209, 212, 217, 223

HST: Hubble Space Telescope. 209

HTTP: HyperText Transfer Protocol. 226

IA: intrinsic alignments of galaxy shapes. 156, 158, 174

IAU: International Astronomical Union. 144

IDAC: Independent Data Access Center. v, vi, 18, 19, 49

Image Reduction and Analysis Facility: a collection of software written at the National Optical Astronomy Observatory (now NOIRLab) geared towards the reduction of astronomical images in pixel array form.. 242

IMBH: Intermediate Mass Black Hole. 82, 83

IMF: Initial Mass Function. 40, 94, 98–100, 177

Independent Data Access Center: Externally supported and administered versions of the DAC to serve the full, or a limited subset of, the LSST data products and/or software to authorized users.. v, 242

IR: infrared. 53, 56, 57, 104, 127, 148, 166, 167, 195

IRAF: Image Reduction and Analysis Facility. 104, 148

ISIS: Interactive Spectral Interpretation System, https://space.mit.edu/cxc/isis/. 104

ISO: Interstellar Object. 145–147

JPL: Jet Propulsion Laboratory (DE ephemerides). 142

JWST: James Webb Space Telescope (formerly known as NGST). 84, 148

K2: NASA mission that provides precise photometric data from numerous target fields in the ecliptic.. 117

KiDS: Kilo-Degree Survey. 156, 159, 168–171, 175

kSZ: kinetic Sunyaev-Zeldovich effect. 166

LAMOST: Large Sky Area Multi-Object Fibre Spectroscopic Telescope, also known as the Guo Shoujing Telescope. 88

LC: Light Curve. 117

LCDM: Λ Cold Dark Matter; cosmological model. 156

LEs: Light Echoes. 104–106

LF: luminosity function. 97

LG: Local Group. 91

LIGO: Laser Interferometer Gravitational-Wave Observatory. 222, 226

LINCC: LSST Interdisciplinary Network for Collaboration and Computing. v, vi, 1–3, 18, 22

LISA: Laser Interferometer Space Antenna. 108

LMC: Large Magellanic Cloud. 87, 89

LRG: Luminous Red Galaxies. 41, 170–172

LSB: Low Surface Brightness. 41–47, 195, 198

LSST: Legacy Survey of Space and Time (formerly Large Synoptic Survey Telescope). v, 1–3, 9, 34–36, 40, 41, 43, 45–47, 52–55, 59–69, 71, 73, 75, 76, 79, 80, 82, 87–94, 96, 97, 102, 105, 106, 108–111, 113, 114, 116–121, 123, 126, 127, 129, 132–146, 148–150, 153, 154, 156, 158, 162, 163, 166–168, 170–174, 189, 191–194, 200, 204, 207–210, 213–215, 222, 224–227, 230

LSST Corporation: An Arizona 501(c)3 not-for-profit corporation formed in 2003 for the purpose of designing, constructing, and operating the LSST System. During

Opportunity: The degree of exposure to an event that might happen to the benefit of a program, project, or other activity. It is described by a combination of the probability that the opportunity event will occur and the consequence of the extent of gain from the occurrence, or impact. There are two levels of opportunities. At the macro level, a project itself is the manifestation of the pursuit of an opportunity. At the element level, tactical opportunities exist, whereby certain events, if realized, provide a cost or schedule savings to the project or increase technical performance. 62, 247

Pan-STARRS: Panoramic Survey Telescope and Rapid Response System. 56, 59, 73, 121, 127, 133, 143, 150, 223, 224

parquet: see Apache Parquet. 231

PASP: Publications of the Astronomical Society of the Pacific. 144

passband: The window of wavelength or the energy range admitted by an optical system; specifically the transmission as a function of wavelength or energy. Typically the passband is limited by a filter. The width of the passband may be characterized in a variety of ways, including the width of the half-power points of the transmission curve, or by the equivalent width of a filter with 100% transmission within the passband, and zero elsewhere. 103, 106

PCA: Principal Component Analysis. 187

PDF: Probability Density Function. 212, 213

photo-z: photometric redshift. 44, 64, 158, 167, 170, 171, 173, 174, 212–215, 219, 220, 225

photometric redshift: Often abbreviated to photo-z, this is an estimate of the true redshift (of a galaxy) determined from multi-band photometry. Generally determined from a fit of source colors to grid of model SEDs with redshift. 4, 40, 44, 65, 79, 80, 159, 244

pipeline: A configured sequence of software tasks (Stages) to process data and generate data products. Example: Association Pipeline. 61, 70, 106, 132, 138, 145–147, 149, 150, 163, 165, 171, 200, 210

postage stamp: Image cutouts that are 30x30 arcseconds, centered on an Object, and included in every Alert. 4, 46, 82, 203, 205

PS1-MDS: PS1 Medium Deep Survey. 63

PSD: power spectral density. 77

PSF: Point Spread Function. 109, 137, 138, 140, 157, 195, 206

PTF: Palomar Transient Factory. 73, 223

Qserv: LSST's distributed parallel database. This database system is used for collecting, storing, and serving LSST Data Release Catalogs and Project metadata, and is part of the Software Stack. 97

R1: Doctoral Universities – Very high research activity. 233

RA: Right Ascension. 62, 115, 149

RAIL: Redshift Assessment Infrastructure Layers, https://github.com/LSSTDESC/RAIL. 20, 174, 220, 221

Release: Publication of a new version of a document, software, or data product. Depending on context, releases may require approval from Project- or DM-level change control boards, and then form part of the formal project baseline. 79, 115, 192, 203, 228, 240

RESSPECT: Recommendation System for Spectroscopic Follow-up. 162, 165

RGB: Red Giant Branch. 87

RMS: Root-Mean-Square. 127

RNADE: Real-valued Neural Autoregressive Distribution Estimation. 57

RNN: Recurrent Neural Network. 69

ROSAT: Röntgensatellit X-ray telescope. 57

RRab: RRL subgroup of fundamental-mode pulsators, most common and display the steep rises in brightness typical of RRL. 126

RRc: RRL subgroup with shorter periods and more sinusoidal variation. These are the less common population of RRL. 126

RRd: RRL subgroup of double mode pulsars and are the most rare RRL. 126

RRL: RR Lyrae stars. 126–128

RSP: Rubin Science Platform. 20, 49, 51, 109, 192, 199, 207, 208, 211, 231

RTA: Real Time Analysis. 204

SACC: Save All Correlations and Covariances. 174

SAGA: Satellites Around Galactic Analogs (Survery). 43

SB: Surface Brightness. 45, 46

SC: Science Collaboration. v, 1–3, 18–20

Science Collaboration: An autonomous body of scientists interested in a particular area of science enabled by the LSST dataset, which through precursor studies, simulations, and algorithm development lays the groundwork for the large-scale science projects the LSST will enable. In addition to preparing their members to take full advantage of LSST early in its operations phase, the science collaborations have helped to define the system's science requirements, refine and promote the science case, and quality check design and development work. v, 20, 151, 240, 245

Science Pipelines: The library of software components and the algorithms and processing pipelines assembled from them that are being developed by DM to generate science-ready data products from LSST images. The Pipelines may be executed at scale as part of LSST Prompt or Data Release processing, or pieces of them may be used in a standalone mode or executed through the LSST Science Platform. The Science Pipelines are one component of the LSST Software Stack. 156, 157

Science Platform: A set of integrated web applications and services deployed at the LSST Data Access Centers (DACs) through which the scientific community will access, visualize, and perform next-to-the-data analysis of the LSST data products. 109, 187–189, 209

SSI: Synthetic Source Injection. 158
SSO: Solar System Object. 227
STEM: Science, Technology, Engineering and Math. 12
Stripe 82: A 2.5° wide equatorial band of sky covering roughly 300 square degrees that was observed repeatedly in 5 passbands during the course of the SDSS, In part for calibration purposes. 222
Structure Function: measure of variance of observations separated in time. 76, 246
survey footprint: The portion of the sky covered by data from an astronomical survey, e.g., the main wide-fast-deep LSST 10-year survey, the LSST deep drilling fields, or the Science Validation data taken during commissioning. Sometimes represented by Boolean maps or other summary statistics in an all-sky representation, e.g., the IVOA MOC standard. 90
SVOM: Space Variable Objects Monitor. 62
Synthetic Source Injection: injecting fake objects onto images to test the detection and measurement process. 158, 247

TAP: Table Access Protocol. 203, 204
TDE: Tidal Disruption Event. 40, 56, 64, 65, 75, 192
TDEs: Tidal Disruption Events. 9, 55, 64, 65, 186
TESS: Transiting Exoplanet Survey Satellite. 120, 121, 130, 224
TGAS: Tycho-Gaia Astrometric Solution. 98, 99
THOR: Tracklet-less Heliocentric Orbit Recovery, an algorithm described in Moeyens et al. (2021). 132–135
TNO: trans-Neptunian object. 137
TNS: Transient Name Server. 65, 224, 226
TOM: Target and Observation Manager. 65, 115, 151, 226, 227
ToO: Target of Opportunity. 62, 63
TOPCAT: Tool for OPerations on Catalogues And Tables. 118
tracklet: Links between unassociated DIASources within one night to identify moving objects. 132, 133
transient: A transient source is one that has been detected on a difference image, but has not been associated with either an astronomical object or a solar system body. vii, 3, 4, 40, 55–57, 59–61, 65, 67, 102, 118, 136, 192, 193, 201–206, 209, 219, 224, 228, 240
tSZ: thermal Sunyaev-Zeldovich effect. 166

UDF: User Defined Function. 126, 128
UNIONS: Ultraviolet Near- Infrared Optical Northern Survey. 171
UT: Universal Time. 104, 208
UV: Ultraviolet. 53, 117
UW: University of Washington. 133

Validation: A process of confirming that the delivered system will provide its desired functionality; overall, a validation process includes the evaluation, integration, and test activities carried out at the system level to ensure that the final developed system satisfies the intent and performance of that system in operations. 163

VISTA: Visible and Infrared Survey Telescope for Astronomy. 74

VLA: Very Large Array (NRAO). 74, 148, 248

VLASS: The Very Large Array Sky Survey carried out by VLA. 74

W3C: World Wide Web Consortium. 226

WCS: World Coordinate System. 229

WFD: Wide Fast Deep. 158, 174

WISE: Wide-field Survey Explorer. 74, 77, 78, 102, 127, 167

WL: Weak gravitational Lens cosmic shear. 201

World Coordinate System: a mapping from image pixel coordinates to physical coordinates; in the case of images the mapping is to sky coordinates, generally in an equatorial (RA, Dec) system. The WCS is expressed in FITS file extensions as a collection of header keyword=value pairs (basically, the values of parameters for a selected functional representation of the mapping) that are specified in the FITS Standard. 229, 248

WTTS: Weak-lined T Tauri stars. 116

XMM: ESA X-ray Multi-mirror Mission. 80, 84

XRISM: X-ray Imaging and Spectroscopy Mission. 84

YSO: Young Stellar Object. 102, 116

zBEAMS: Extension of BEAMS light curve classification method to include redshift (z) information. 160, 161

ZTF: Zwicky Transient Facility. 35, 60, 61, 63, 66, 68–71, 73, 77, 83, 114, 117, 121, 123, 127, 130, 133, 140, 204, 223, 224, 226