# Snowmass2021 - Letter of Interest

# *Algorithmic Advances for Processing Data from Cosmological Surveys*

**Thematic Areas:** (check all that apply □/■)
■ (CompF1) Experimental Algorithm Parallelization
□ (CompF2) Theoretical Calculations and Simulation
■ (CompF3) Machine Learning
□ (CompF4) Storage and processing resource access (Facility and Infrastructure R&D)
□ (CompF5) End user analysis
□ (CompF6) Quantum computing
□ (CompF7) Reinterpretation and long-term preservation of data and code

**Contact Information:**
Andrew Connolly (University of Washington) [ajc@astro.washington.edu]
Collaboration: The Vera C. Rubin Observatory LSST Dark Energy Science Collaboration (DESC)

**Authors:** Yusra AlSayyad (Princeton University), Andrew Connolly (University of Washington), Katrin Heitmann (Argonne National Laboratory), Robert H. Lupton (Princeton University), Peter Melchior (Princeton University), Rachel Mandelbaum (Carnegie Mellon University)

**Abstract:**

Cosmology has come a long way in the last few decades, from providing a broad understanding of the evolution and content of the Universe to precision measurements that have helped establish the "Standard Model of Cosmology". The data being collected by ongoing and upcoming surveys will enable us to sharpen the constraints even more and guide us in our quest to understand the nature of the dark Universe. However, in order to fully exploit the promise of the new data sets, continuous advances in the algorithms deployed to process the data will be essential. Many challenges lie ahead: sheer size is an obvious hurdle, but the unprecedented data quality and the emergence of new computational architectures will also require continuous improvements of the processing algorithms. New directions for algorithm development will need to be explored and the performance of already existing algorithms will have to be improved. Finally, support and academic recognition for the development of high-performance and high-precision algorithms will be needed to establish a community of physicists that can fully exploit the next generation of cosmological surveys.

# 1 Introduction

Enormous resources have been poured into developing new experiments and instruments that can survey the universe over many decades of the electromagnetic spectrum. Ground-based surveys such as the Rubin Observatory's Legacy Survey of Space and Time (LSST) will repeatedly image the sky at visible wavelengths generating petabytes of images over a 10 year period. Space-based telescopes such as the Nancy Grace Roman Survey Telescope (RST) or the Euclid satellite will generate high resolution infrared images of billions of stars and galaxies. At millimeter wavelengths, CMB-S4 will map the primordial fluctuations and polarization patterns of the Cosmic Microwave Background. The science these experiments will deliver is diverse, including measurements of the properties of dark energy to an accuracy 10x better than today (showing whether the acceleration of the universe can be explained by modifications to general relativity or by the presence of an energy density that arose in the very early universe); detection of primordial gravitational waves as a probe of inflationary physics; and mapping the evolution of the clustering and distribution of matter within the universe.

The scientific impact of these experiments will, in large part, be driven by our ability to process and analyze the data they generate (often in almost real time). Continued development and enhancement of advanced and novel algorithms will be critical to their success. In this LOI, we describe the impact that such an investment in algorithmic development over the lifetime of the Cosmic Frontiers missions would enable and a possible path towards realizing these opportunities.

# 2 Challenges and Opportunities

Algorithmic development in image processing has dramatically increased the scientific returns from many previous surveys. For example, for the case of the Sloan Digital Sky Survey (SDSS), it was only a year after first light that an accurate model for the point-spread-function (PSF) of the SDSS telescope and its variation across the focal plane was developed. The advent of algorithms that circumvented the need for accurate modeling of PSFs in extremely crowded fields enabled large-scale image difference pipelines and the emergence of time domain astrophysics. Global photometric calibration techniques developed for the SDSS, five years after the start of the survey, are now adopted throughout current imaging surveys (including the Dark Energy Survey and the LSST). Even seven years after first light for the SDSS, systematics introduced via sky subtraction around bright galaxies that biased the photometry of background sources were being uncovered and corrected, enabling more robust measurement of galaxy-galaxy lensing on small scales.

SDSS's work on algorithms to deblend multicolor galaxy images allowed reliable photometry of Luminous Red Galaxies (LRGs), enabling the detection of Baryon Acoustic Oscillations. The requirements of deeper modern surveys such as Rubin's LSST require the development of more sophisticated approaches. Two concrete example are 1) Scarlet[1], a newly developed multi-band source deblender, and 2) Metacalibration[2] for shear measurement. All such processing algorithms need to provably deliver unbiased and accurate results and at the same time be feasible to use within available computing resources.

Common software frameworks could be deployed across multiple communities within HEP even as individual experiments modify and adapt their systems to specific applications. As noted in a number of reports on the future of computing, such an approach would not preclude developing alternative or competing algorithms where scientific requirements drive the need for different solutions, even for the same instrument. It could instead simplify the continued development and optimization of algorithms and applications over the life of an experiment, accounting for changes in hardware, computational architectures, and new statistical methodologies and improving not just the precision of Cosmic Frontier science but also the breadth of science that these experiments can deliver.

As described in the Petabytes to Science whitepaper[3], many of the advances in computational, algorithmic, and statistical techniques, whether developed within the Cosmic Frontiers community or beyond, are demonstrated as proofs of concept on small-scale data sets. These new algorithms do not perform uniformly

well when faced with the realities of a petabyte imaging survey. Expertise in instrumentation is needed to convert these algorithms into applications robust to new edge cases that arise from the instrument or unique survey strategy. Performance at-scale requires software engineering expertise that is not easily found within the academic community. Furthermore, algorithms from external disciplines often make assumptions that are challenging to translate into use cases for Cosmic Frontiers experiments, obscuring their advantages or disadvantages. Here again, researchers with expertise in software engineering and the characteristics of an experiment can play a critical role. The lack of this expertise can limit the adoption and sharing of new methodologies across domains.

New computational architectures, e.g. the emergence of cloud-based architectures or high-performance computers based on GPUs, require education and training if they are to be used effectively. Transitioning the community to work with these systems as an integral part of the computing infrastructure will require sustained support. For many cases, where data resources are stored or how they are provisioned does not matter to an end user as long as data products are delivered efficiently and at little to no cost. Designing algorithms that can effectively use these resources does, however, require a detailed knowledge and experience with the underlying system.

Education and training of a community with research interests in algorithms and both the experience with data and the expertise to know what tools are available and how to work with them becomes critical. The current lack of academic recognition and a career path (formal or informal), which would lead to the presence of experts in the universities and national labs, creates a disincentive to develop the skills that are needed for an era rich in data.

# 3 Summary

Sustainable software development, to support the lifecycle of software from prototyping ideas to the delivery of supported and maintained packages, remains a serious challenge. This is particularly the case when the underlying computational architecture changes over the lifetime of a survey (e.g. the emergence of GPU or many-core hardware with limited memory per core).

Support for the training of physicists (both early career and more experienced researchers) would build a community of scientists familiar with the complexities of survey data and the best practices for software design and development. Recognition of the intellectual merit of novel data analysis approaches, and ensuring continued support for technical and scientific experts between missions and experiments, in universities and national labs, would provide a career trajectory that might reduce the loss of this expertise to industry.

## References


[1] https://pmelchior.github.io/scarlet/index.html

[2] https://github.com/esheldon/ngmix/wiki/Metacalibration

[3] A.E. Bauer, E.C. Bellm, A.S. Bolton et al., arxiv:1905.05116 [astro-ph.IM]