Snowmass2021 - Letter of Interest

Software and Statistics for Discovery in Cosmic Frontiers Experiments

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)
Brian Nord (Fermilab/UChicago) [nord@fnal.gov]
Collaboration: The Vera C. Rubin Observatory LSST Dark Energy Science Collaboration (DESC)

Authors:
Simon Birrer (Stanford University), Andrew Connolly (University of Washington), Leanne P. Guy (Vera C. Rubin Observatory/NOIRLab), Katrin Heitmann (Argonne National Laboratory), Brian Nord (Fermilab)

Abstract:
Thanks to investments by the DOE, NSF, NASA, and international agencies, new generations of experiments, instruments, and surveys are being constructed to probe the evolution and formation of the universe. The LSST (Rubin Observatory's Legacy Survey of Space and Time) will quantify the nature of dark energy by repeatedly surveying the near and distant universe; Euclid and RST (the Nancy Grace Roman Space Telescope) will map the distribution of dark matter at higher resolution than ever before; and CMB-S4 (Next Generation CMB Experiment) will probe inflationary physics by measuring the fluctuations and polarization of the microwave background. The science we will achieve from these experiments will be driven by how well we can sift through the unprecedented volumes of complex data to search for fundamental truths. Therefore, software will be a key instrument for exploring the universe. To support the science from these experiments, we will need to implement research programs that can support the full life cycle of algorithmic and software development (from the initial ideas and development of statistical techniques to the creation and maintenance of robust software frameworks). Creation of scalable software infrastructure running in the cloud (public or academic), workshops to bring together researchers from different fields, seed-funding competitions, incubators and fellowships to support the development of new methodologies from the broad scientific community, and an educational program to transform the physics communities into data-aware and data-literate scientists capable of exploiting large-scale data streams will all need to be part of this program.

1 Introduction

Astrophysics is undergoing a major paradigm shift, continuing its transformation into a data-rich field where multi-petabyte spatio-temporal data sets are commonplace. This data explosion is causing the very nature of astronomical investigations to change.The advances during this decade are expected to be weighted towards exploratory studies examining whole data sets: performing large-scale classification (including using machine learning techniques), clustering analyses, searching for exceptional outliers, or measuring faint statistical signals. The shift is driven by the needs of the scientific questions of the day. For some – such as the nature of dark energy – utilization of all available data and improved statistical treatments are the only way to make progress. For others, such as time series data, detailed insights into the variable and transient universe will only come if we can learn the properties of entire populations of variable sources and not just individual objects.

This LOI is based on the discussions and recommendations of the Petabytes to Science workshop held in Boston (November 2019)[1]. We describe the opportunities and challenges that will arise from large, complex data sets generated by current and planned Cosmic Frontier experiments. By creating and supporting programs for the development of algorithmic techniques (together with the statistical frameworks that underpin them), the construction of robust visualization and analysis platforms, and the implementation of formal and informal training programs we can develop a collaborative, sustainable, and ethical approach to statistical astrophysics that will drive much of the science in the next decade.

2 Recommendations

Our objectives with these recommendations are to enable any undergraduate student, graduate student, postdoc, or senior researcher to easily apply their tools and methodologies to data sets at any scale, to spur the creation of new computational and statistical frameworks needed to analyze a new generation of data sets, to support the creation of astrophysics data, to make the tools to analyze them accessible to the whole community, and to jump start the training (at any stage of a researcher's career) of a generation of data literate scientists. Projects of this scope could not easily be carried out by individual universities but with coordination across agencies, national labs, universities and experiments a program to benefit physics and astrophysics as a whole could be brought into being.

- **Create sustainable funding models for data intensive science:** Statistical methodologies and their validation requires sustained development and funding. Programs that can support all phases of this development cycle (workshops to define the problem sets and possible solutions, resources to prototype applications, the creation of robust software development best practices, validation frameworks for the application of statistical methods to large data sets, and support for software and data curation and maintenance) are required throughout the lifespan of Cosmic Frontier experiments.

- **Understanding correlations and outliers in large data sets:** Discovery in astrophysical surveys requires that we can detect subtle signatures of physics within complex data sets and identify those data that don't match these correlations (i.e. outliers). High dimensionality increases the complexity of this process as the number of sources required to characterize any correlation will often grow exponentially with the number of measured features or attributes. Funding the development of statistical methodologies and their implementation within a software framework would open an area of data exploration to build on current model-fitting methodologies.

- **Develop cloud analysis environments for science:** The cloud provides a natural platform for sharing of data and analysis tools across and within experiments. For example, in the Cosmic Frontier, easy sharing of data between experiments would simplify the correlation of observations across multiple wavelengths. Given that different wavelengths probe different physical environments the correlations

between these data can lead to new probes of cosmology (e.g. the cross-correlation of CMB and galaxy distributions to measure the integrated Sachs Wolfe effect). Easy sharing of tools within a common software environment would also enable tools and analyses developed within an experiment to migrate to other experiment.

- **Create formal and informal educational programs to train physicists in statistical methodologies and software engineering:** The education and training of students must keep up with the developments and advances in statistics, machine learning, and software engineering. The emergence of data-intensive science has not been tracked by the educational and training programs available to graduate students, postdoctoral fellows, and senior personnel. Creating common curricula and material that can be shared across institutions, providing specialized or advanced training programs aimed at more advanced researchers, developing workshops and schools that extend beyond a single course, and extending the training of physics and astrophysics students to include machine learning, visualization, computational infrastructure, and scalable analytics would provide an educational foundation for HEP research programs. By targeting programs to fill the gaps in education that are not currently provided by universities this could enable data science education for a broad range of institutions (e.g. providing expertise that may not be available at smaller institutions)

- **Promote and support open software and open distribution of methodologies**: The development of open source software and the open publication of algorithms within journals such as the Journal of Open Source Software has increased the participation of researchers in software engineering over the last decade. As an example, the astropy project has had over 340 unique contributors to its code base in the nine years it has been under development. It maintains one of the most actively used code bases in astrophysics yet until 2019 there was less than one funded FTE working on the project. Funding to support the continued development and maintenance of open source programs is critical to their longevity. Recognition of the work open source contributors make, through prizes, awards, postdoctoral fellowships, and career advancement, could increase the number of researchers who can engage in open source programs and, thereby, accelerate the adoption of new and innovative methodologies within the HEP community.

- **Evaluate the ethical and societal implications of advanced statistical techniques:** Machine learning and statistical applications are already having profound ethical and societal implications — even if they are developed solely to address specific scientific questions within HEP. The assessment of ethical implications for algorithm development and application should become a standard step in the research process. Additionally, ethics should become a standard educational element during the training of HEP researchers. Finally, HEP researchers should partner with ethicists and science and technology studies professionals in the formulation of policies, processes, and educational materials.

References

[1] Amanda E. Bauer, Eric C. Bellm, Adam S. Bolton, Surajit Chaudhuri, A. J. Connolly, Kelle L. Cruz, Vandana Desai, Alex Drlica-Wagner, Frossie Economou, Niall Gaffney, J. Kavelaars, J. Kinney, Ting S. Li, B. Lundgren, R. Margutti, G. Narayan, B. Nord, Dara J. Norman, W. O'Mullane, S. Padhi, J. E. G. Peek, C. Schafer, Megan E. Schwamb, Arfon M. Smith, Erik J. Tollerud, Anne-Marie Weijmans and Alexander S. Szalay, 2020, arXiv 1905.05116