Peter Nugent, Division Deputy for Scientific Engagement
Computational Research Division, LBNL
Adjunct Professor of Astronomy, UC Berkeley
M.S. 50B-4206 - 1 Cyclotron Road - Berkeley, CA, 94720-8139
Phone:(510) 486-6942 - Fax:(510) 486-4300 - Cell:(925) 451-4001
E-mail: penugent@LBL.gov - Web: http://www.lbl.gov/~nugent

Role and point of view in the big data innovation:

I began my research life as a theorist and have since transitioned to observational astrophysics. All of the major work I have carried out in my career has involved high-performance computing (HPC). I have been able to follow a variety of scientific fields, and the effect that changes in HPC architectures and tools has had on them over the past two decades due to my joint position at the National Energy Research Scientific Computing Center (NERSC) and at the Computation Research Division at LBL. Since 2008 I have focused almost exclusively on challenges in data analysis, specifically mining extremely large astrophysical data sets and confronting these directly with even greater sized simulation data. This RFI offers the opportunity to create a community from which one could leverage help to advance three major themes I see as crucial in the coming decade for data analysis: taking full advantage of new HPC resources (multi-core/many-core, BurstBuffer, high-speed networking, etc.); novel techniques for confronting experimental data with simulations; and devising courses at both the undergraduate and graduate level to educate our students on the best ways to address these problems in the coming decade.

Extreme Data Analysis in Cosmology:

In recent years astrophysics has undergone a renaissance, transforming from a data-starved to a data-driven science. A new generation of experiments — including Planck, BOSS, DES, DESI, Euclid, WFIRST and LSST — will gather massive data sets that will provide more than an order of magnitude improvement in our understanding of cosmology and the evolution of the universe. Their analysis requires leading-edge high performance computing resources and novel techniques to handle the multiple PB's of data generated throughout these surveys. Furthermore, interpreting these observations is impossible without a modeling and simulation effort that will generate orders of magnitude more "simulation" data — used to directly understand and constrain systematic uncertainties in these experiments and to determine the covariance matrix of the data. As we enter this era of precision cosmology a thorough propagation of errors on measurements in both the experiments and simulations becomes essential. This is especially true in modern cosmological models where many parameters and measurements are partially degenerate, and such degeneracies can lead to important shifts in the cosmological parameters one is trying to measure.

Data Co-Design is a term that refers to a computer system design process where scientific problem requirements influence architecture design and technology, and constraints inform formulation and design of algorithms and software. To ensure that future architectures are well-suited for data target applications and that major data scientific problems can take advantage of the emerging computer architectures, major ongoing research and development centers of computational science need to be formally engaged in the hardware, software, numerical methods, algorithms, and applications co-design process. An example of this is engaging next-generation HPC centers in their engineering efforts on future big-iron machines they will deploy for use by the general scientific community. For example, the software development effort for BurstBuffer (a tier of solid-state storage systems to absorb application I/O requests) in order to more readily handle the analysis of ~PB datasets in cosmology as outlined above. As there will be many science applications that could benefit from such an effort, NITRD could aid in the expansion of the co-design process to more fields across several agencies.

Simulation-Based inference for cosmology:

The DOE, NASA and NSF have invested significantly in large-scale Dark Energy experiments (the Dark Energy Survey – DES, the Large Synoptic Survey Telescope – LSST and the Wide Field infrared Telescope - WFIRST) where observations of Type Ia supernovae (SNe Ia) play a central role. They have also been a driving force behind the computational science of SN Ia physics as well as large-scale simulations of the structure of the universe. While large-scale simulations have a long history of using simulations *directly* in the analysis of observational data, the SN field does not. Today, SN cosmologists still rely on fully empirical techniques to chase Dark Energy. These methods are derived from data that suffer important selection biases and calibration issues, using parameterizations that may be only qualitatively justified after the fact. The gulf between simulation and observation is especially disappointing given that many SN Ia systematic errors affecting Dark Energy measurements should be best (or may only be) addressed through simulation. This is an area that many experimental and observational programs lack as well. An effort to promote the infrastructure to bridge the gap between the simulation and modeling effort with data analysis can pay huge benefits.

Furthermore, we need to take a serious look at how these agencies share these large data sets and analysis tools. Currently it is everyone for themselves and whatever home-brew method they come up with. This creates a lot of hurdles in the joint analysis of data sets that span these agencies.

Education:

In the spring of 2013 I taught a course at Cal entitled, High Performance Computing for Astrophysicists. The goal of the course was to provide an introduction to Unix and the working environment on HPC systems. Students were given accounts at NERSC in order to gain experience running a variety of current parallel codes in astrophysics and handling both the resultant large datasets generated by these simulations as well as observational datasets through NERSC's Science Gateway Nodes. I think the time is ripe to work on creating a more standard set of courses in our university system, applicable to a variety of scientific domains, in which one can expose students to a variety of techniques in data analysis that highlight the power of novel algorithms, HPC architectures and the latest in I/O, databases, etc. Getting NITRD to promote this educational effort across the various agencies would create huge benefits in both the near and long-term future for data science.