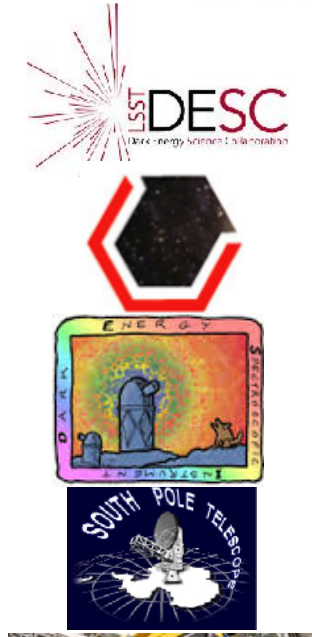
SciDAC: Accelerating HEP Science — Inference and Machine Learning at Extreme Scales

Focus Areas:

- **Cosmology:** Unique arena for advanced stats/ML applications — big data, big compute, large-scale inverse problems
- **'Stats/ML at Scale':** Need to speed up methods by ***many orders of magnitude*** to enable dealing with datasets and science requirements in the multi-PB to EB era
- **Accuracy:** Many problems in a regime where statistical errors are subdominant — need to understand how to deal with modeling/ mitigating systematics



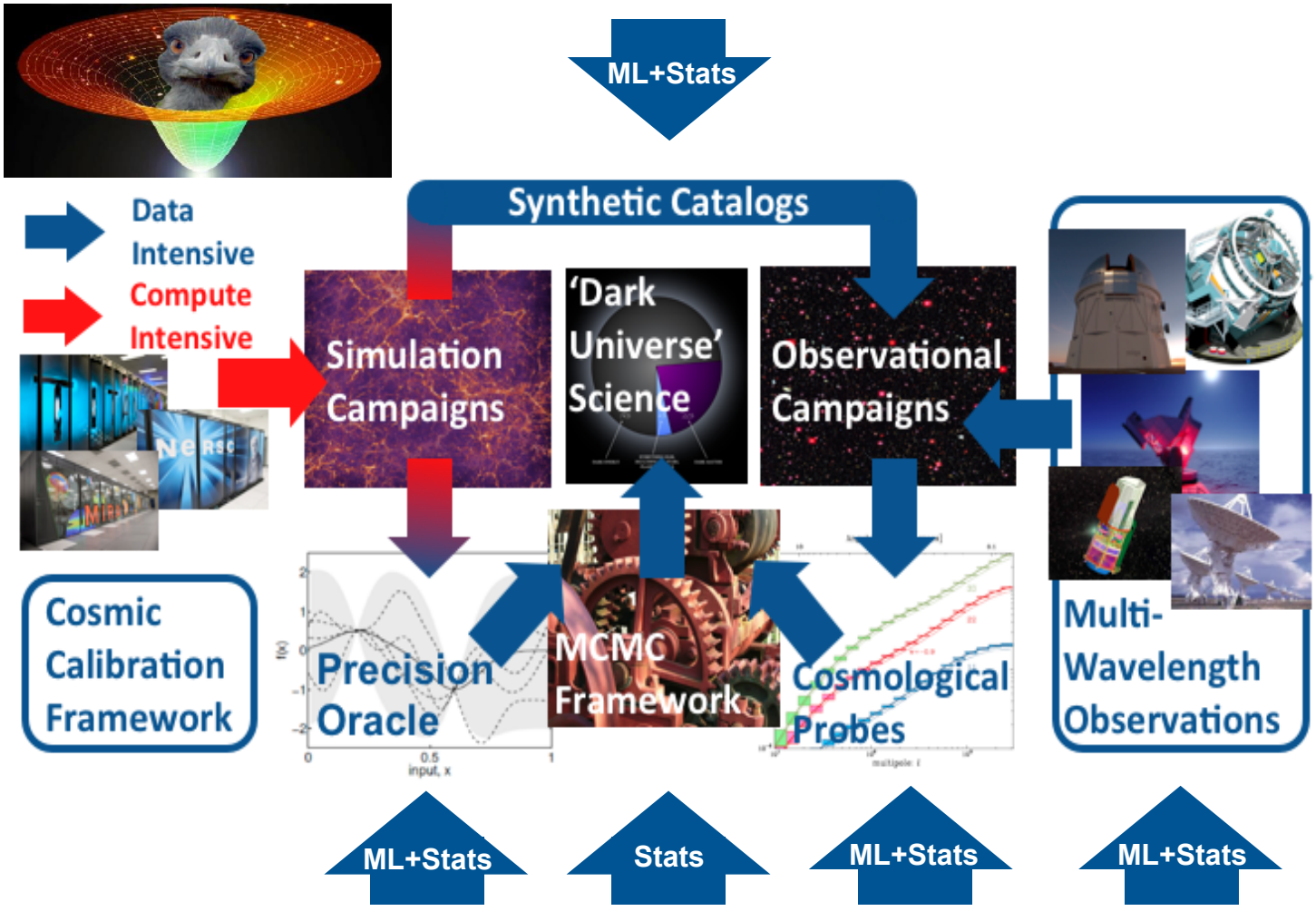
Team: P. Balaprakash, M. Binois, S. Habib (PI), K. Heitmann (Argonne PI), E. Kovacs, N. Ramachandra, S. Wild (Argonne); A. Fadikar, R. Gramacy, D. Higdon (Va Tech PI) (Va Tech); E. Lawrence (LANL, Dep. PI); Y. Lin, A. Slosar (BNL PI), S. Yoo (BNL); Z. Lukic (LBNL PI), D. Morozov (LBNL)

Website: https://press3.mcs.anl.gov/cpac/projects/scidac

Software portal: http://www.hep.anl.gov/cosmology/CosmicEmu/emu.html

Workshop: Argonne, Sep 24-25, 2018 – "Advanced Statistics Meets Machine Learning" (https://indico.fnal.gov/event/18318/overview)

Science with Surveys as an Inverse Problem: Extreme-Scale Computing meets Statistics and Machine Learning



Cosmological scientific inference process showing forward modeling and systematic error exploration/control loop

- Use of HPC resources as high-fidelity, large data-volume sources for state-of-the-art data-intensive statistical and machine learning (ML) methods
- Need to speed up the forward modeling process, deal with 'curse of dimensionality' in the inverse problem
- How to control errors if the modeling and measurement error PDFs are uncertain?

New Techniques for Photometric Redshift Estimation

Scientific Achievement

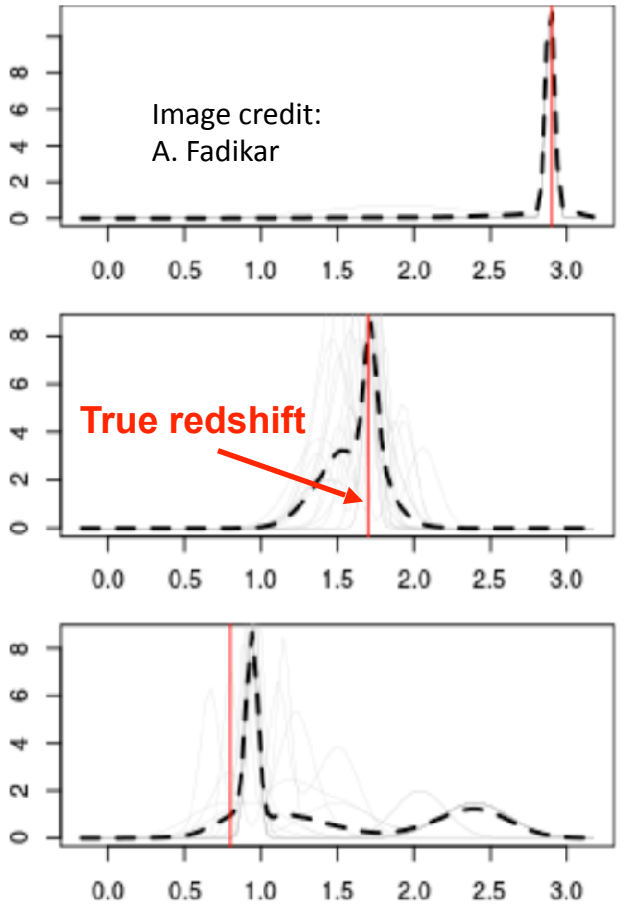
Estimation of galaxy redshift distribution using photometric information, morphology, and spatial correlations; application to LSST

Significance and Impact

Characterization and reduction of photometric redshift estimation errors essential for success of imaging surveys

Research Details

- Large synthetic dataset based on realistic templates for spectral energy distributions (SEDs) of different galaxy types
- Machine learning techniques for classification (hidden space variables), use of mixture models; Bayesian learning for posterior PDFs
- Early results show great promise for photometric redshift estimation applications and error mitigation

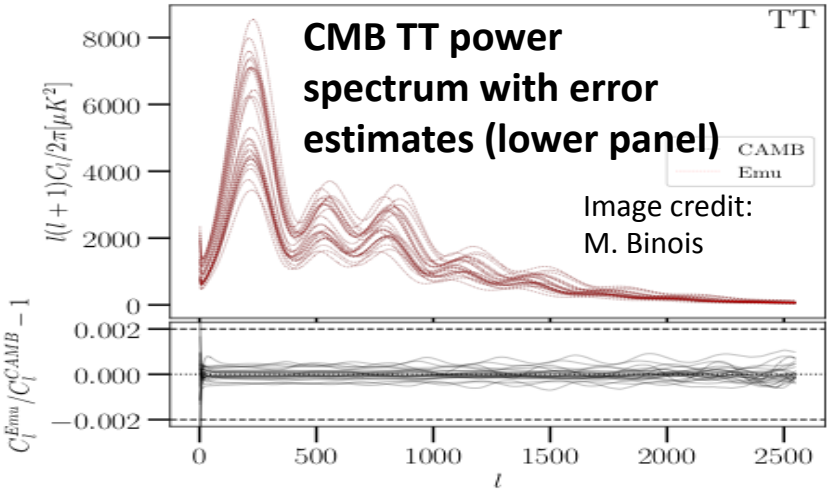


Multi-Gaussian Process approach to obtain estimated redshift PDFs and comparisons to training set values (red vertical lines)

Precision Emulation of CMB Power Spectra

Scientific Achievement

Fast, accurate prediction of cosmic microwave background (CMB) variables (~2000X speedup with 0.2% errors over the desired dynamic range

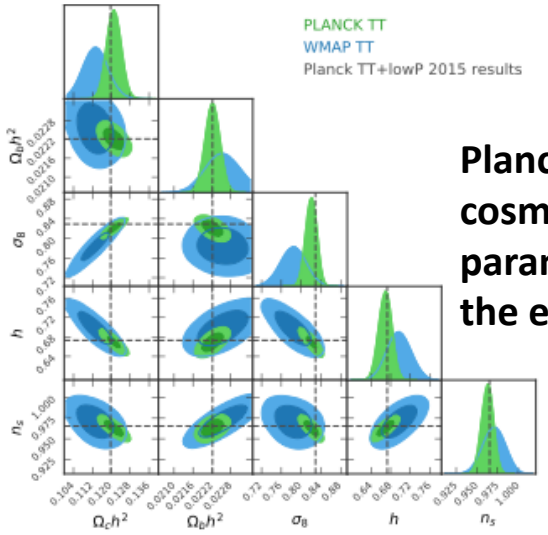


Significance and Impact

Predictions/forecasts for next-generation CMB surveys (CMB-S4), analysis of current-generation data (ACTPol, Planck, SPT-3G, —)

Research Details

- Large training/validation dataset generated using the CAMB code with symmetric Latin hypercube sampling
- Dimensional reduction via unsupervised learning (comparison of variational autoencoders and PCA)
- Non-parametric, Gaussian Process-based interpolation; error estimates via MCMC



Planck vs WMAP cosmological parameters using the emulator

Image credit:
N. Ramachandra

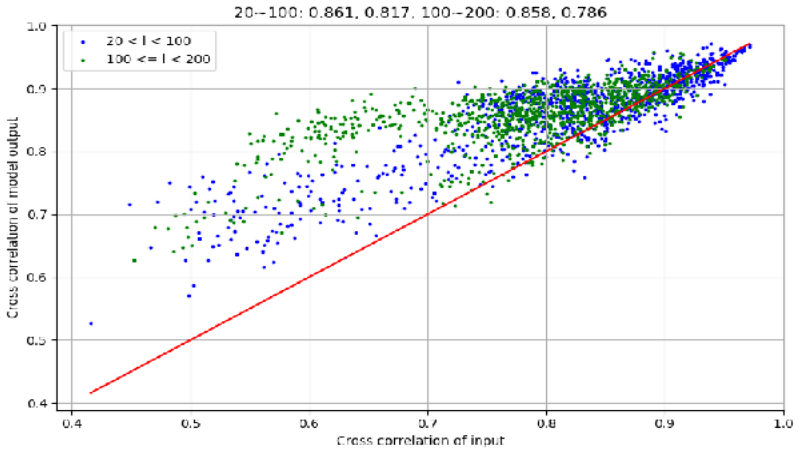
Neural Network Prediction of CMB Dust Foreground

Scientific Achievement

Significant improvement in prediction for dust foreground using convolutional neural networks and Galactic neutral hydrogen data

Significance and Impact

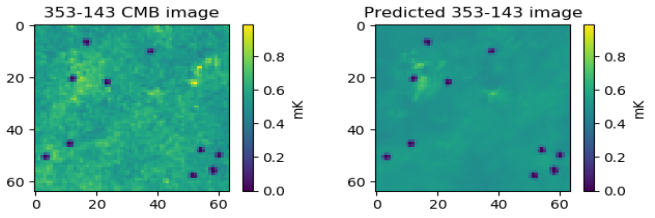
Current work uses intensity data, the next generation will focus on polarization to help with optimal field selection and data analysis for a small aperture CMB-S4 experiment



Improvement in cross-correlation coefficient with target map compared to naive total intensity map; above red-line indicates improvement, below indicates deterioration (Green is for ~1 deg scales, blue is for ~10 deg scales)

Research Details

- Input data are 50 velocity slices in galactic neutral hydrogen as traced by the 21 cm line
- Output data are the difference in Planck 353GHz and 143GHz data which is dominated by the dust signal
- Trained on southern galactic hemisphere, validated/tested on the northern galactic hemisphere
- Optimal linear model gives negligible improvement: neural net is picking up nontrivial information



Example prediction: Planck difference map (left) and model prediction (right); black circles are point-source mask

Images credit: G. Zhang

Basic Emulation for Ly-alpha Forest Statistics

Scientific Achievement

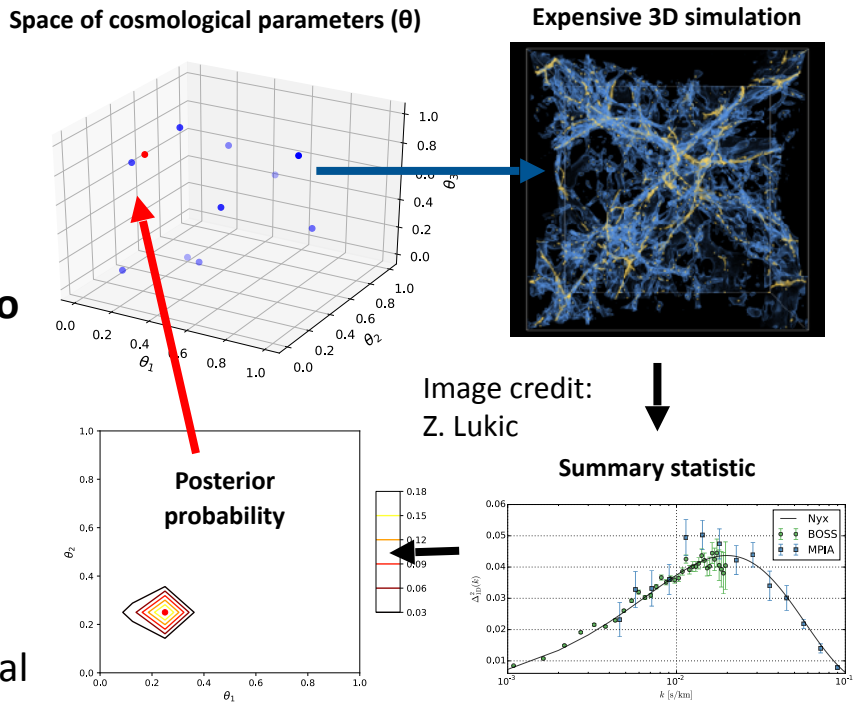
HPC framework to infer cosmological and thermal parameters using Ly-alpha power spectrum and selected computational model runs

Significance and Impact

Ly-alpha forest observations are the main window into structure formation at high redshifts (2<z<5) and a sensitive probe of non-CDM cosmologies. P(k) emulation is necessary for recovery of cosmological parameters from observations

Research Details

- Automated system for iteratively running cosmological simulations and analysis tasks on HPC systems
- New iterative method to determine most informative points in parameter space for running the next batch of simulations
- Multiple ways to do GP emulation of vector summary statistics, i.e., exploring ways of combining k- and cosmological parameter dependence of emulated P(k)



Inferring cosmological parameters in a 3-parameter test case: Simulations produce matter configurations that depend on cosmological parameters; Ly-alpha analysis produces outputs comparable to sky survey measurements. Combining predictions and measurements we infer "current best" parameter probabilities as well as the "promising" points for the next set of simulations in our iterative procedure.

Image Classification/Regression for Strong Lensing

Scientific Achievement

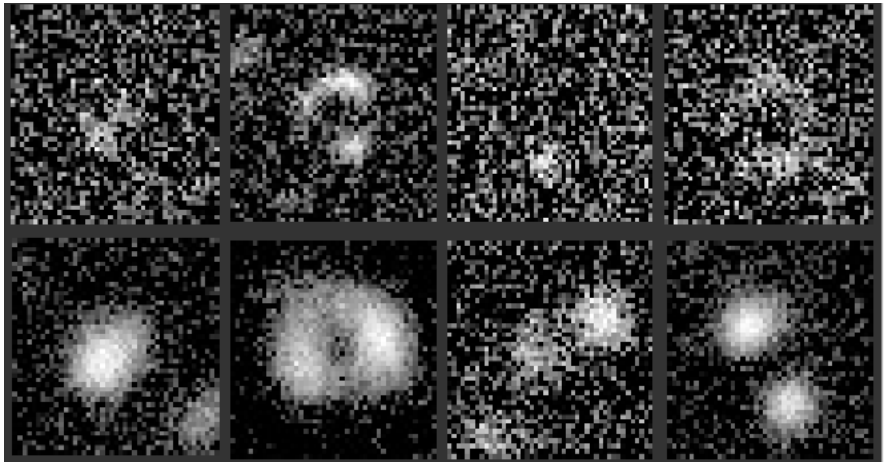
Fast (10 microseconds/image), robust (80-90% accuracy) classification of strongly lensed background galaxies

Significance and Impact

LSST will have tens of billions of objects with ~100K strongly lensed sources — automated source detection/filtering is essential

Research Details

- Large synthetic dataset based on full ray tracing algorithm with 1) model halo mass distribution as lenses and 2) halos from cosmological simulations, realistic telescope properties; single and stacked images
- Deep CNN classification/regression
- GANs for fast generation of new images



Single and stacked noisy lensed training images for LSST; a companion set of non-lensed images is not shown

Image credit:
N. Li

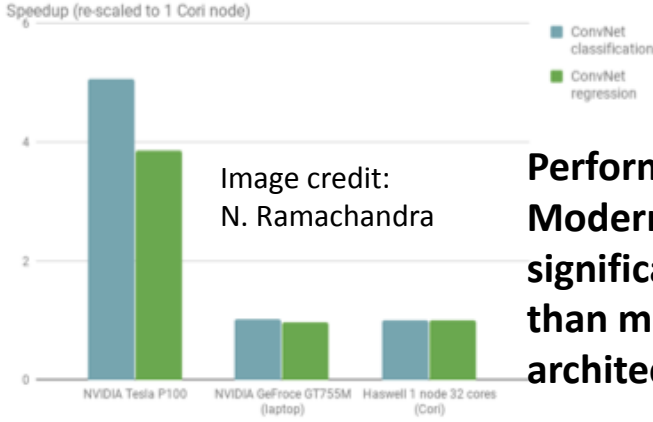


Image credit:
N. Ramachandra

Performance: Modern GPUs are significantly faster than manycore architectures

Future Work

- **Emulation Landscape:**
  - Extend work on summary statistics to problems with significantly higher dimensionality, *O*(10) to *O*(100)
  - Multi-fidelity emulation
  - Develop new methods for applications to likelihood-free scenarios (e.g., semi-analytic galaxy modeling)
  - Fast generation of multiple realizations of 'raw' sky data; develop techniques for ensuring dynamic consistency (causality vs. correlations)
- **Image Applications:** Image cross-validation, source de-blending algorithms, application to calibration studies
- **ML/DL Methods on HPC Platforms:** Work on scaling up ML and statistical methods on HPC platforms with GPU acceleration (e.g., Cooley@ALCF, Summit@OLCF)
- **Stats meets ML:** Improve methods by incorporating model information into 'black box' techniques; incorporate optimization methods into Bayesian calibration, many other topics —