LA-UR-20-27968
Approved for public release; distribution is unlimited.

Title: Foresight: Analysis That Matters for Data Reduction

Author(s): Grosset, Andre Vincent Pascal
Biwer, Christopher Michael
Pulido, Jesus J.
Mohan, Arvind Thanam
Biswas, Ayan
Patchett, John M.
Turton, Terece
Rogers, David Honegger
Livescu, Daniel
Ahrens, James Paul

Intended for: SuperComputing 2020, 2020-11-15 (Atlanta, Georgia, United States)

Issued: 2020-10-08

Foresight: Analysis That Matters for Data Reduction

Pascal Grosset, Christopher M. Biwer, Jesus Pulido, Arvind T. Mohan, Ayan Biswas, John Patchett, Terece L. Turton, David H. Rogers, Daniel Livescu, James Ahrens

Los Alamos NATIONAL LABORATORY EST.1943

SC20 Everywhere we are | more than hpc.

ECP EXASCALE COMPUTING PROJECT

ExaSky: Computing the Sky at Extreme Scales

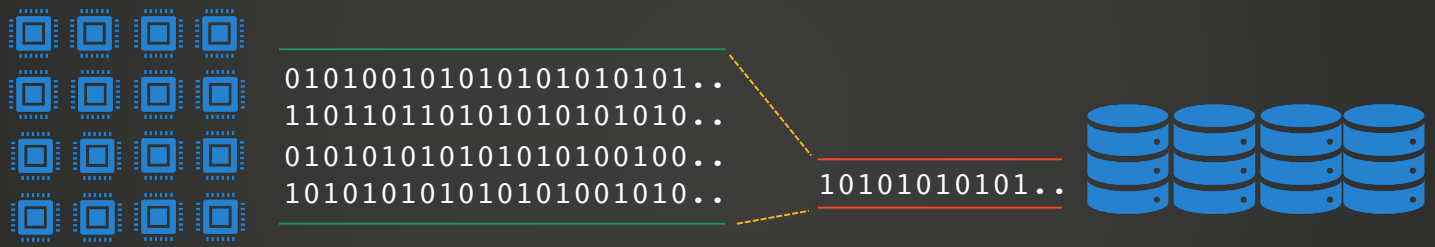
November 19, 2020

NNSA National Nuclear Security Administration

Managed by Triad National Security, LLC for the U.S. Department of Energy's NNSA

Exascale Computing and Data Explosion

- Simulations are generating massive amounts of data, overwhelming I/O capabilities

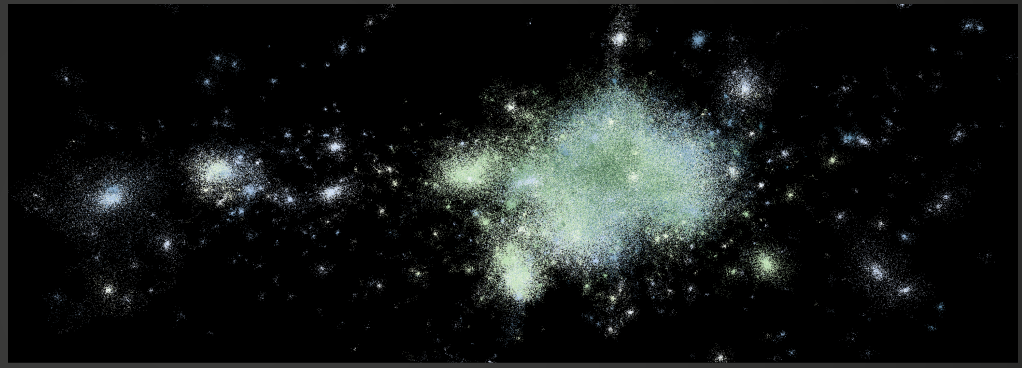


- Compression:
  - Lossless: Ideal but only about 2X compression ratios for scientific data
  - Lossy: How much data can we afford to lose?

Data Reduction Questions: Precision needed

Precision required?

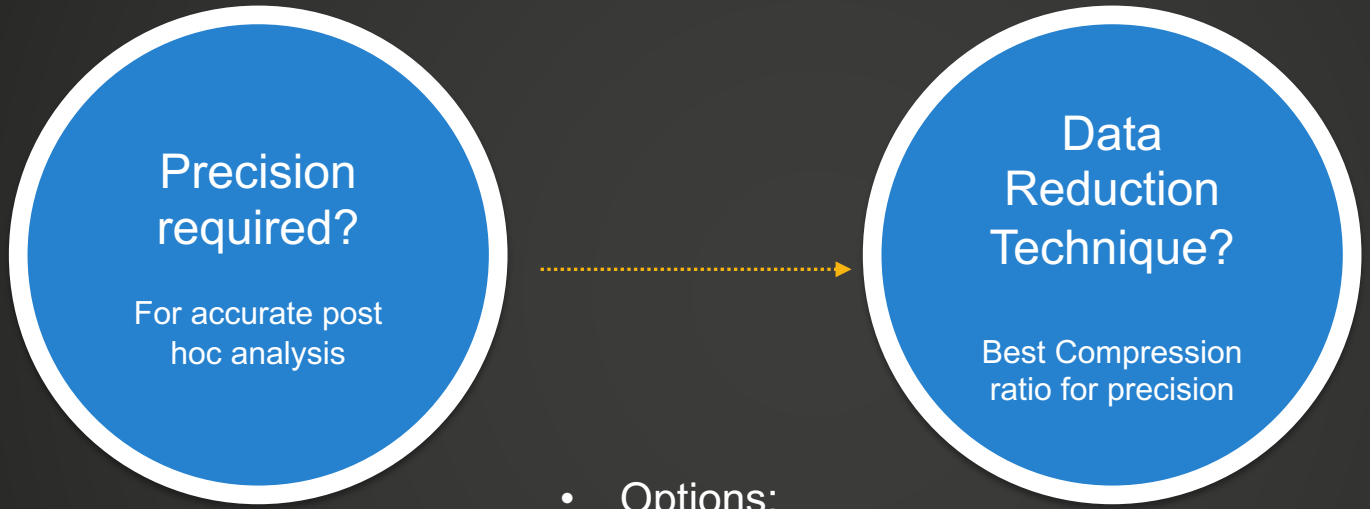
For accurate post hoc analysis



- Determined by the scientist:
  - By how much will the center of a halo move after particle position are compression?
  - Will the mass of a halo be the same?
  - By how much will the power spectrum deviate?

~~PSNR~~

Data Reduction Questions: Precision needed

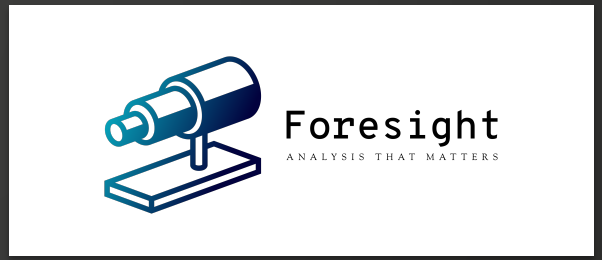


- Options:
  - “Traditional algorithms” e.g. SZ, ZFP,…
  - Auto-encoders (machine learning)
  - Sampling
- Each compression methods has a large parameter space to sweep through

Contribution

- Foresight:
  - A framework that enables us to **compare** different data reduction schemes
  - Evaluate how they **impact** post-hoc analysis

- Stages:
  - Data Reduction
  - Analysis
  - Visualizing Comparison



- Run **analysis that matters** to scientist for data reduction
- Bring **Verification and Validation** to compression

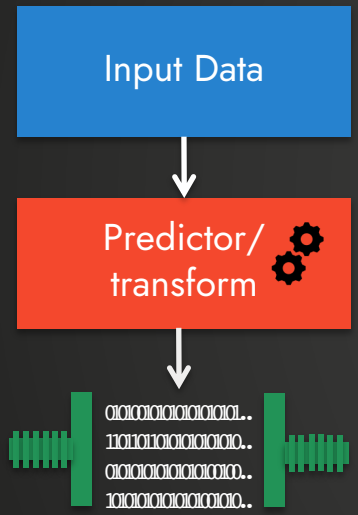
Outline

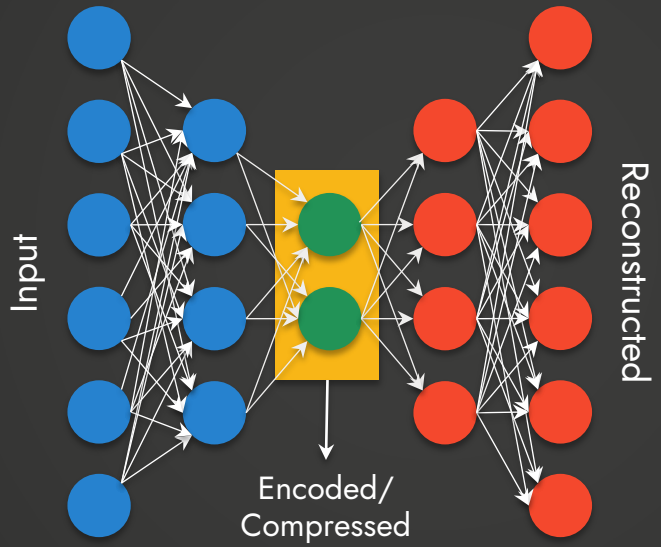
- Data Explosion and Data Reduction
- Contribution
- Outline
- Related Works
- Architecture of Foresight
- Use Cases
  - Cosmology: Compression
  - Turbulence: Machine learning
  - Asteroid Impact: Sampling vs “Traditional” Compression
- Conclusion

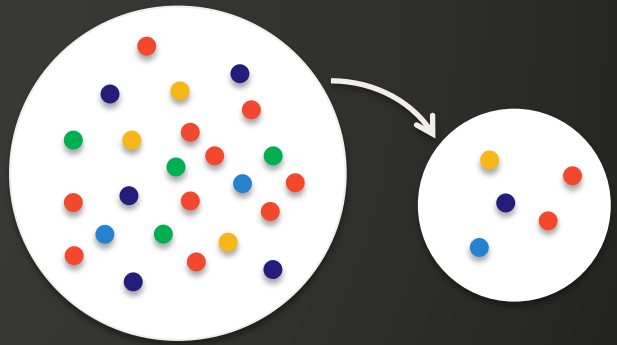
Related Works: Compression Evaluation

- Compression Metrics:
  - “Standard Metrics” – compression ratio, throughput, …
- Custom Tools:
  - Squash
  - Z-Checker
  - Libpressio
- Custom validation by scientists

Related Works: Methods





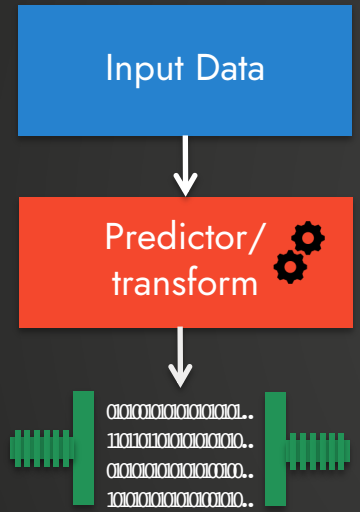


"Traditional Compressor"

Autoencoders

Sampling

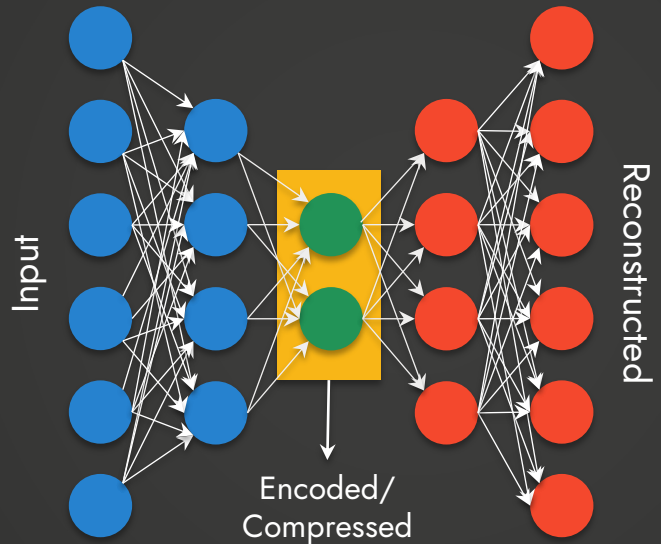
Related Works: Methods



"Traditional Compressor"

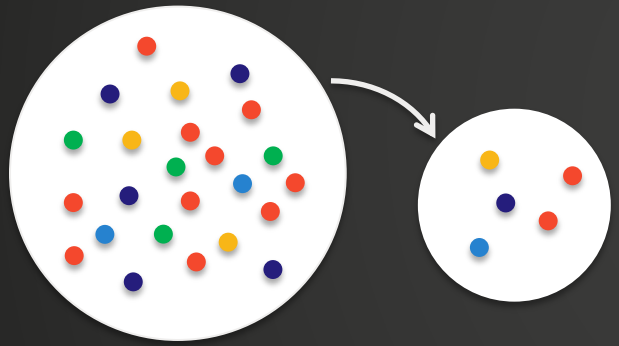
- Common Options
  - SZ, ZFP, ISABELA, …
- Methods:
  - Predictor
  - Transform

Related Works: Methods



Autoencoders

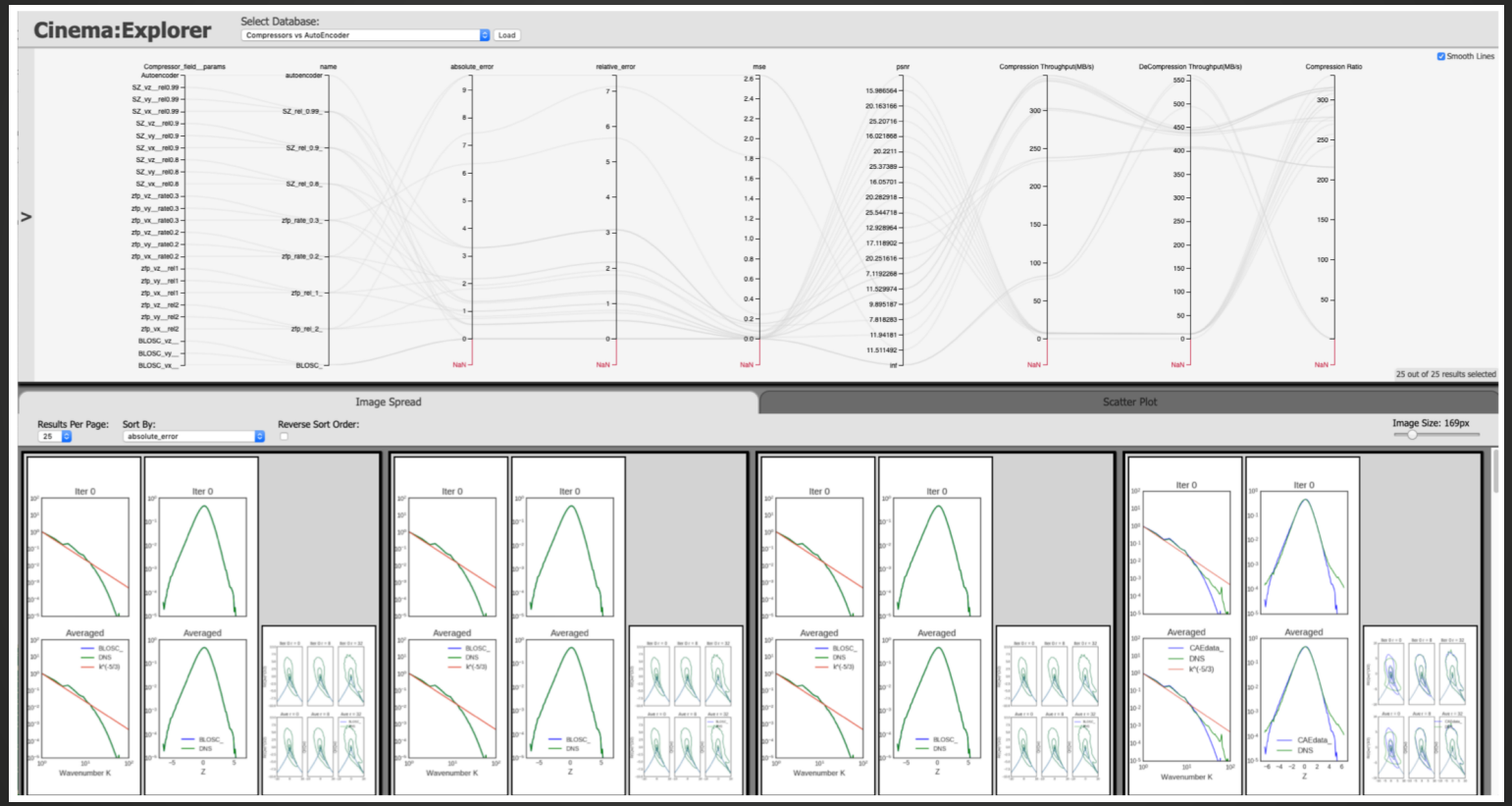
Related Works: Methods



Sampling

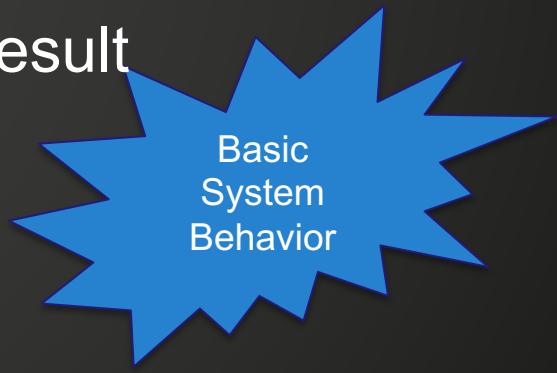
- Common Options
  - Random
  - Regular
  - Histogram-driven

Related Works: Data Presentation



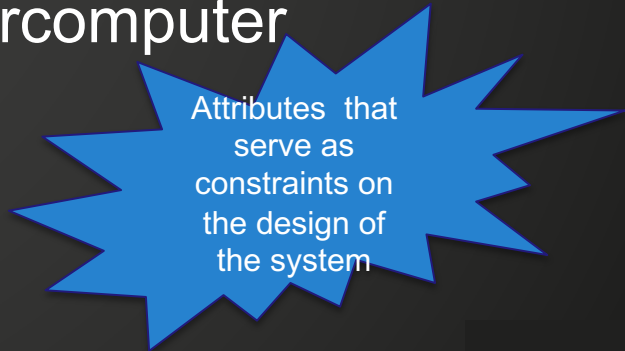
Architecture: Functional Requirements

1. Supports different data reduction techniques
2. Ability to split data on multiple ranks
3. Run post hoc analysis on supercomputers
4. Support visualization and exchange of result

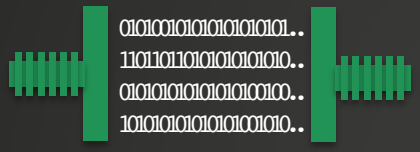


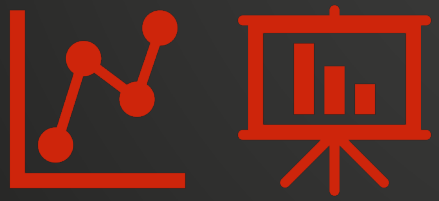
Architecture: Non Functional Requirements

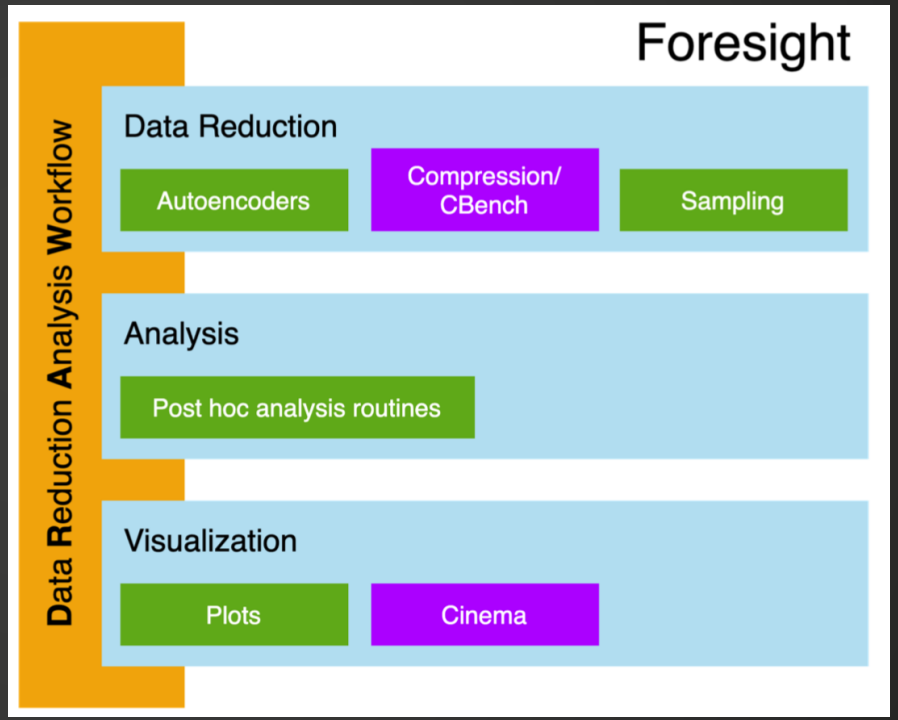
1. Provenance
2. Simple API for specifying data reduction, analysis, and visualization
3. Logging
4. Extensibility: Easy of adding new features
5. Ease of building and running on supercomputer
6. Self-Verification



Foresight: Architecture

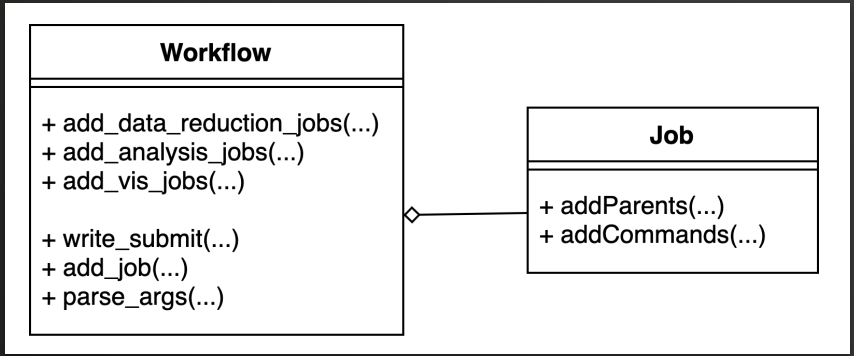


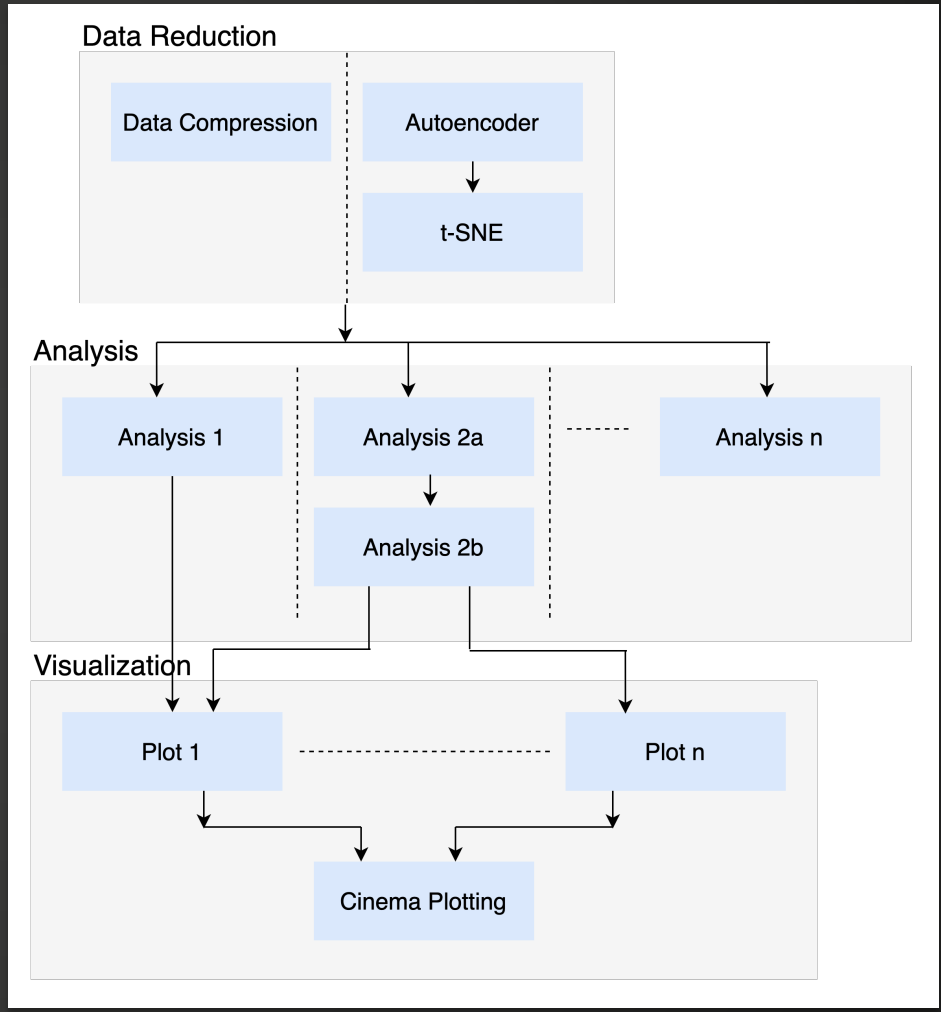




Foresight: DRAW

- Heart of Foresight
  - Python based
  - Reads in configuration JSON file

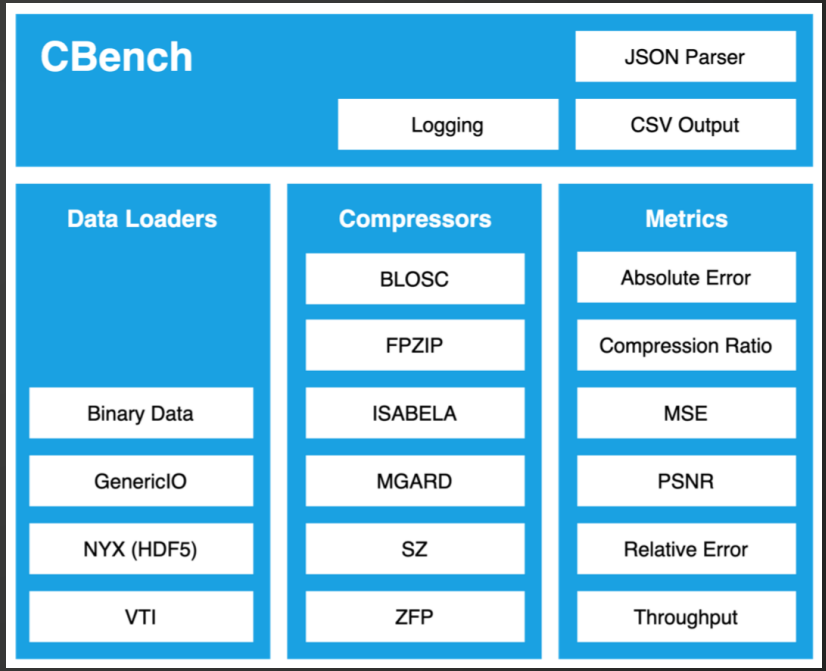




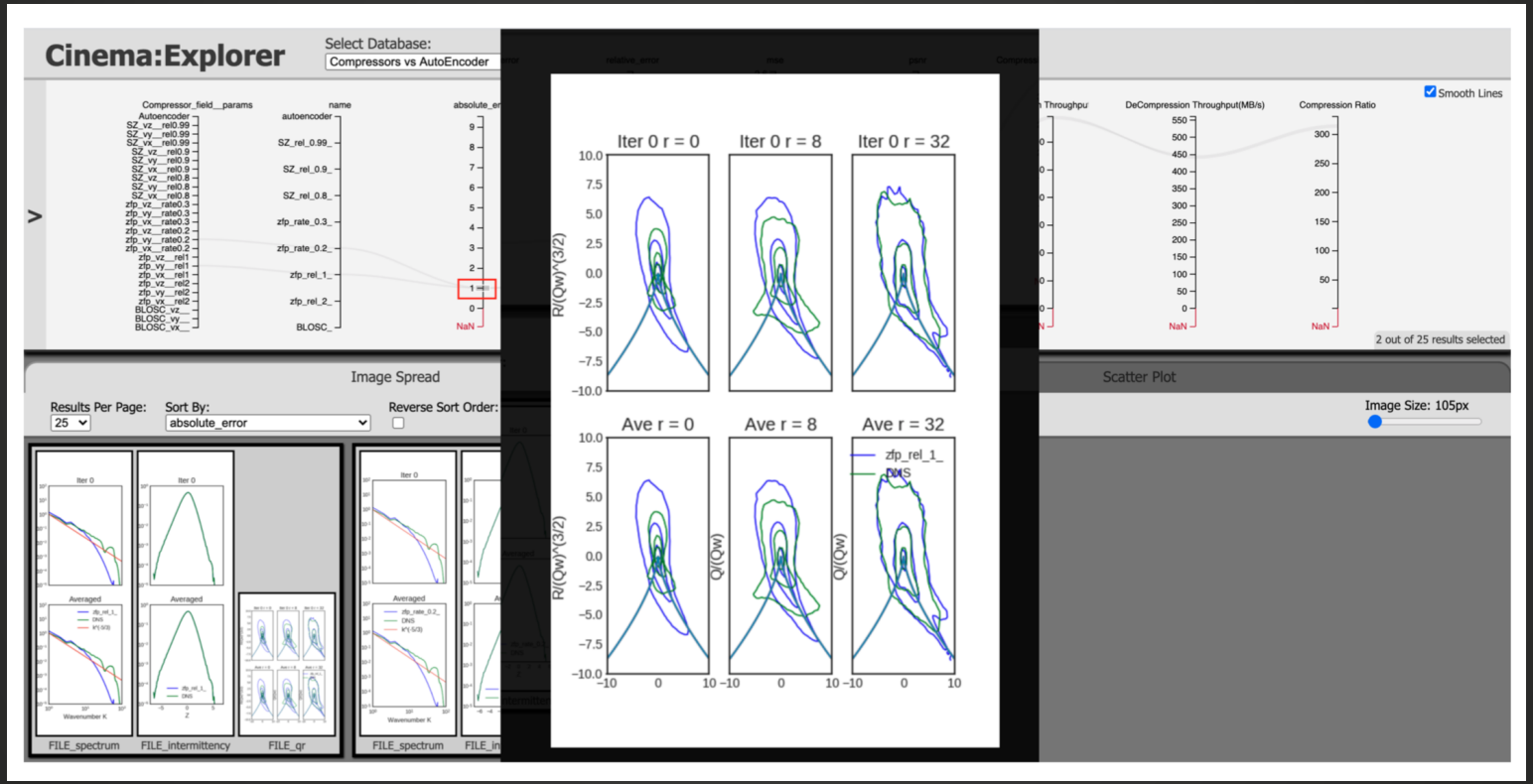
Foresight: CBench

- Characteristics:
  - C++/MPI
  - Abstract Factory design pattern

```
read input JSON file;
initialize reader;
for each timestep do
    for each compressor do
        for each data field to be processed do
            load compressor parameter;
            compress data;
            for each compressor metric to be measured
              do
                compute metric;
            output metrics to disk;
        Save decompressed data to disk;
```



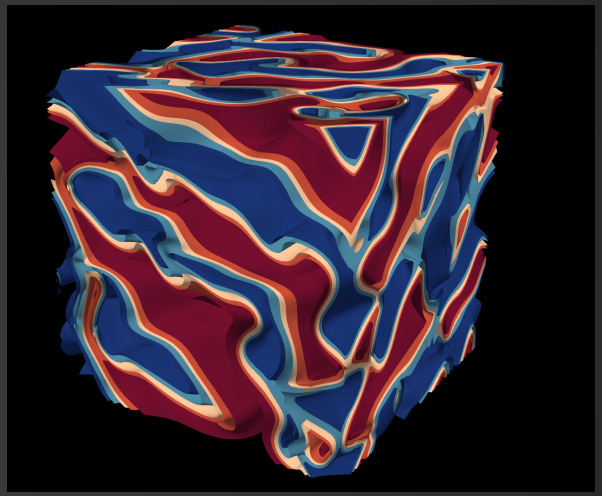
Foresight: Visualization



Turbulence: Machine Learning

3D Direct Numerical Simulation dataset

- Box size of 128^3, with 50 timesteps over 2.5 GB
- Multiscale analysis is used to quantify feature preservation



DNS Turbulence Density Visualization

Data Reduction in Turbulence

- Datasets are extremely large
- Turbulence is highly multiscale in nature
- Compression and modeling of turbulence requires the large inertial-scale features to be accurately retained, while small-scale not much

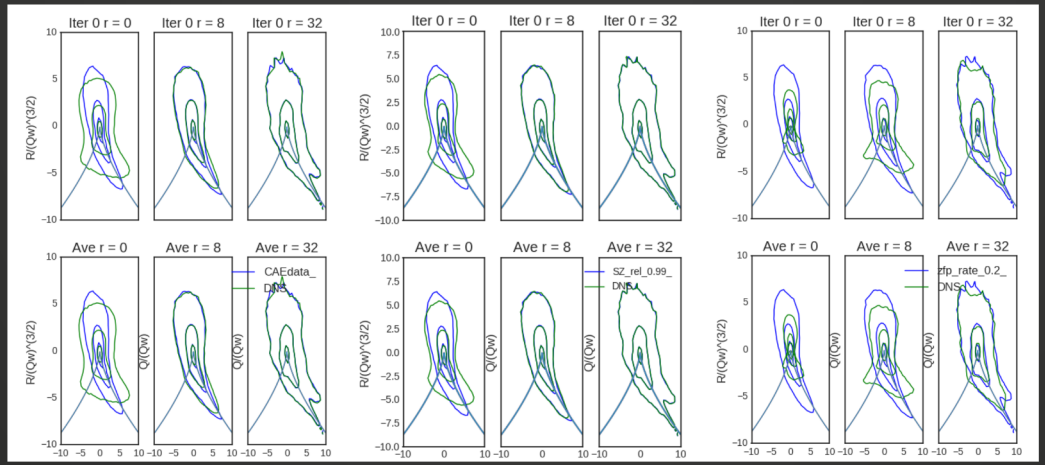
Data Reduction Requirement

- ***CFDNS-A***: Which data compression algorithm will give the best throughput, PSNR, and MSE for a very large compression ratio (e.g., 300X compression), typical of autoencoders?
- ***CFDNS-B***: How do data compression algorithms compare with autoencoders for capturing not just the large scale portion of the energy spectrum, but also the 3D morphology of the flow?

Turbulence: Machine Learning

Data compression Algorithm	Parameters	Variable	Absolute Error	Relative Error	PSNR	MSE	Throughput (MB/s)	Compression Ratio
SZ	relative error 0.99	vx	3.29	3.09	0.08	25.21	7.44	330.71
		vy	0.52	0.52	0	20.16	7.53	307.4
		vz	1.35	1.35	0.02	15.99	7.53	269.19
ZFP	relative error 1.0	vx	9.53	7.45	2.63	9.9	346.11	310.9
		vy	0.99	0.99	0.01	11.53	345.57	312.11
		vz	2.17	2.17	0.15	7.12	340.5	311.8
ZFP	fixed rate 0.2	vx	9.53	7.45	2.63	9.9	338.29	315.08
		vy	0.99	0.99	0.01	11.53	342.25	315.08
		vz	2.17	2.17	0.15	7.12	338.73	315.08

CFDNS-A is addressed by computed logging and computing general statistics native to Foresight.



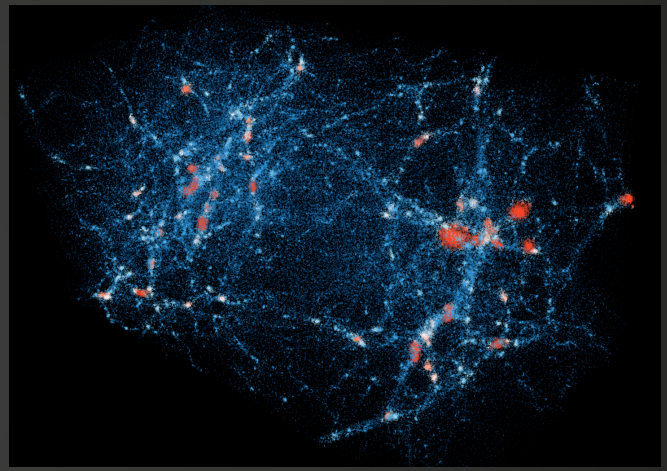
Autoencoder SZ ZFP

A QR-plot, native to turbulence analysis, is used to map behavior and resolves **CFDNS-B**.

Cosmology: Compression

The HACC cosmology dataset
- A large 45GB particle dataset with 256 partitions
- Halo regions are used for features of interest

Data Reduction in Cosmology
- Cosmology is an observational science
- Acceptable data loss within 3 kiloparsec in error
- BLOSC is already in use, but does not provide sufficient compression



Cosmology visualization of a HACC particle dataset, colored by particle density

Data Reduction Requirement

- ***HACC-A***: Which compressor will give the best compression ratio and throughput for a maximum absolute error of 0.003?
- ***HACC-B***: What is the impact of an absolute error of 3 kpc on the power spectrum and halos?
- ***HACC-C***: How much can we increase the compression ratio before the data become unusable?

Cosmology: Compression

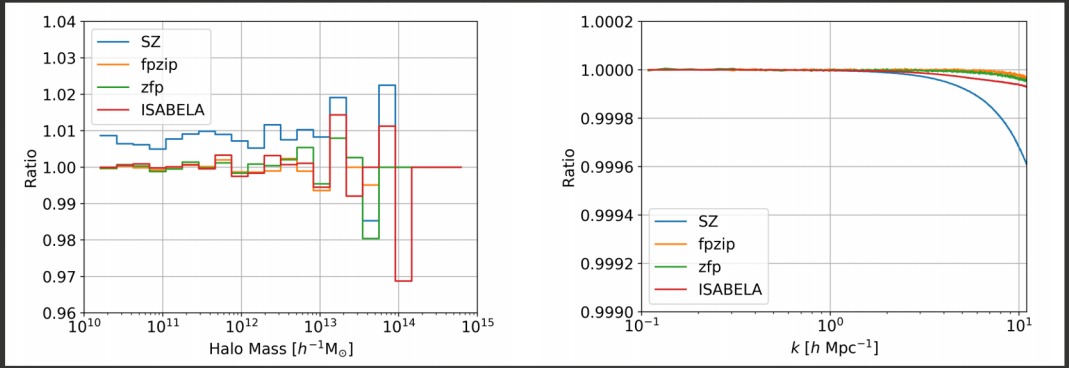
Data compression Algorithm	Parameters	Variable	Absolute Error	PSNR	MSE	Throughput (MB/s)	Compression Ratio
fpzip	Truncate: 26 bits	x	0.000961	115.72	1.71e-07	14.48	3.18
		y	0.000961	115.52	1.79e-07	14.06	3.12
		z	0.000961	115.52	1.78e-07	13.95	3.04
ISABELA	Window Size: 2048 Tolerance: 0.001 Coefficients: 50	x	0.00256	112.85	3.40e-07	1.24	2.21
		y	0.00256	111.96	4.17e-07	1.03	2.09
		z	0.00256	110.69	5.60e-07	0.85	1.49
SZ	Absolute Error Bound: 0.003	x	0.00299	103.42	2.99e-06	77.52	**4.93**
		y	0.00299	103.40	2.99e-06	76.89	**4.78**
		z	0.00299	103.41	2.99e-06	73.95	**4.60**
ZFP	Absolute Error Bound: 0.007	x	0.00291	111.76	4.37e-07	46.41	2.07
		y	0.00243	111.761	4.37e-07	47.55	2.06
		z	0.00242	111.76	4.37e-07	46.63	2.04

A compression table showing the best set of parameters to reach a maximum absolute error target of 0.003 for each position field is used to resolve **HACC-A.**

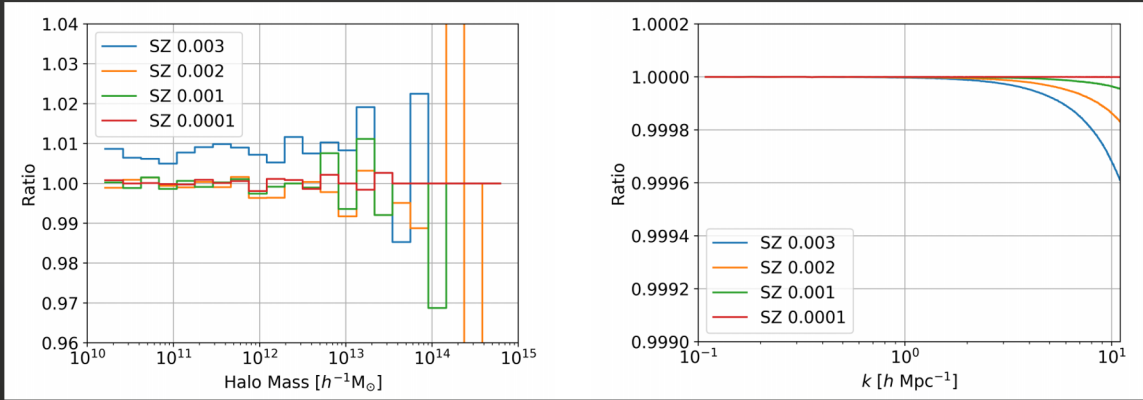
	fpzip	ISABELA	ZFP	SZ
Compression Time (s)	1.61	26.85	0.22	0.28
Decompression Time (s)	1.40	1.77	0.24	0.33

Additional timings for the analysis performed for **HACC-A**.

Cosmology: Compression



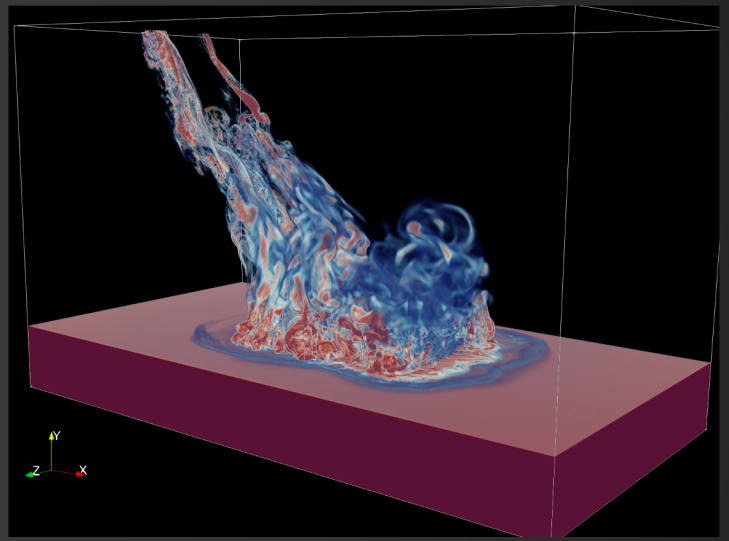
Derivative Halo Analysis looking at mass distribution and relative power spectra ratio are used to answer **HACC-B** to target a max 3kpc error.



Focusing on SZ, it is noted that error falls well within the error budget to answer **HACC-C.**

Asteroid Impact: Sampling vs “Traditional” Compression

The asteroid impact dataset is a simulation by xRage
- 475 timesteps, regular grid data
- Crater impact at t=99 is used for feature analysis.



Volume visualization of the asteroid impact dataset

Data Reduction in Deep Water Impact

- Massive ensemble dataset that often requires downsampling to be shared
- Lossless methods have been used but do not provide enough reduction
- Crater and spray of water ejected must be preserved for post-hoc analysis

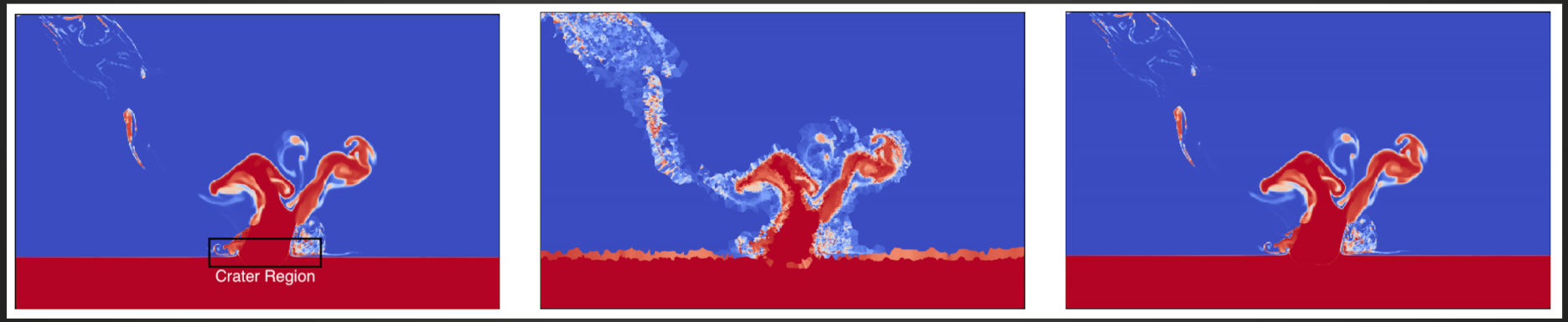
Data Reduction Requirement

- ***xRage-A:*** Which data reduction scheme will give the best compression quality when targeting 50X compression?
- ***xRage-B:*** How does sampling compare to data compression algorithms at the asteroid impact crater?

Asteroid Impact: Sampling vs “Traditional” Compression

Reduction Method	Parameters	Absolute Error	PSNR	MSE	Throughput (MB/s)	Compression Ratio
Regular Sampling	Rate: 0.015	1.0	23.796	0.00417	73.5	48.23
Random Sampling	Rate: 0.015	1.0	23.158	0.00483	72.8	49.94
Histogram Sampling	Rate: 0.010	1.0	19.0459	0.01245	7.93	49.52
SZ	abs 0.0003	0.000300	107.202	1.904e-11	151.576	52.15
ZFP	abs 0.01	0.002262	81.49	7.096e-09	250.806	46.73

General Foresight analysis was performed targeting a 50x compression ratio to answer **XRAGE-A**.



Original | Histogram Sampling | SZ Compression

Asteroid crater region analysis is performed and visualized here, showing comparisons between different methods to address **XRAGE-B**.

Conclusion

- We developed a tool to evaluate the impact of data reduction methods on scientific data by considering impactful analysis that matters
- The framework enables for simple and rapid development of new readers and compressors
- Evaluation metrics from the simulations enable meaningful comparative analysis

https://github.com/lanl/VizAly-Foresight/

Acknowledgements

We would like to thank the National Energy Research Scientific Computing Center (NERSC) for providing access to the Cori supercomputer and LANL for access to the Darwin Supercomputer.

We would also like to thank the HACC team at Argonne National Laboratory for granting us access to cosmology datasets.

This research was supported by the Exascale Computing Project (http://www.exascaleproject.org), a joint project of the U.S. Department of Energy's Office of Science and National Nuclear Security Administration, responsible for delivering a capable exascale ecosystem, including software, applications, and hardware technology, to support the nation's exascale computing imperative.

Thank You!