In Situ Data-Driven Adaptive Sampling for Large-scale Simulation Data Summarization

Ayan Biswas
Los Alamos National Laboratory
ayan@lanl.gov

Soumya Dutta
Los Alamos National Laboratory
sdutta@lanl.gov

Jesus Pulido
Los Alamos National Laboratory
pulido@lanl.gov

James Ahrens
Los Alamos National Laboratory
ahrens@lanl.gov

ABSTRACT

Recent advancements in high-performance computing have enabled scientists to model various scientific phenomena in great detail. However, the analysis and visualization of the output data from such large-scale simulations are posing significant challenges due to their excessive size and disk I/O bottlenecks. One viable solution to this problem is to create a sub-sampled dataset which is able to preserve the important information of the data and also is significantly smaller in size compared to the raw data. Creating an in situ workflow for generating such intelligently sub-sampled datasets is of prime importance for such simulations. In this work, we propose an information-driven data sampling technique and compare it with two well-known sampling methods to demonstrate the superiority of the proposed method. The in situ performance of the proposed method is evaluated by applying it to the Nyx Cosmology simulation. We compare and contrast the performance of these various sampling algorithms and provide a holistic view of all the methods so that the scientists can choose appropriate sampling schemes based on their analysis requirements.

CCS CONCEPTS

• **Mathematics of computing → Statistical paradigms**; • **Human-centered computing → Scientific visualization**; *Visualization techniques*; *Visual analytics*;

ACM Reference Format:
Ayan Biswas, Soumya Dutta, Jesus Pulido, and James Ahrens. 2018. In Situ Data-Driven Adaptive Sampling for Large-scale Simulation Data Summarization. In *Proceedings of ISAV Workshop 2018 (ISAV'18).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

With modern day supercomputers and their immense compute capabilities, large-scale datasets can now be generated in the order of Gigabytes to Petabytes with very high spatial and temporal resolutions. As has been well documented, the disk I/O capabilities of these machines are lagging behind making it impossible to store the full resolution of the data. With the current drive towards the Exascale (10^{18} flops) infrastructures, the disparity between what can be generated and what can be stored has increased significantly. Apart from what can be stored to the disk, another important aspect is what sized data the domain scientists would want to carry with them for post-processing and exploration. With Petabytes to potentially Exabytes of data to explore, post-processing full resolution datasets will soon become prohibitive.

Sub-sampling of datasets for data downsizing has been a popular approach for addressing the aforementioned issues. The application of several well-known in situ sampling methods exist in recent literature, e.g., stratified random sampling [24], bitmap indexing [20, 22], adaptive sampling [17] etc. Although these proposed methods have liberated the burden of post-processing the raw data to some extent, with the Exascale machines on the horizon, very soon we will be needing very low sampling rates (e.g., 1% to 0.1% or less) for the data storage. These existing sampling algorithms are not fully geared towards such requirements as they primarily assume equal importance to all the data values. Since not all regions of a dataset are equally interesting to the scientists, to be more effective in the sampling scenario, novel data-driven sampling methods are required. In [17], although the authors provided an importance-driven data reduction scheme, the importance function was assumed to be known. Instead, for wider applicability of an in situ sampling algorithm, it should be as generic as possible by deriving the importance function from the data itself such that it can preserve the important features at very low sampling rates.

In this work, we propose a new in situ information-driven sampling technique that can be applied across different scientific features and is built upon a generic notion of information theory. The proposed sampling scheme aims at maximizing the information content in the sub-sampled data. Using the basic information theoretic ideas of entropy maximization, our proposed sampling method creates a novel importance metric that generally prioritizes the rare data values of high importance [7]. We compare and contrast our sampling scheme with regular and stratified random sampling schemes to highlight the feature preserving capability of the proposed method along with an in situ performance study demonstrating it's in situ applicability for large-scale simulations.

Our contributions in this work are as follows:

- We propose a novel information-driven sampling scheme for down-sampling regular grid data in situ that preserves the important features of the data.

ISAV '18, November 12, 2018, Dallas, TX, USA

ACM ISBN 978-1-4503-6579-6/18/11…$15.00
https://doi.org/10.1145/3281464.3281467

- We create an in situ framework for the Nyx Cosmology simulations with different sampling methods and show the efficacy of our proposed method.

2 RELATED WORKS

Due to the recent necessity of in situ techniques for large-scale simulation data analysis, several generic in situ capable infrastructures have been developed which add in situ functionality into existing visualization frameworks [10, 13, 14, 21, 23]. Besides directly visualizing the data in situ, some researchers have also developed methods that produce in situ data summaries in various forms. The primary goal of such summary data is to use the reduced data for flexible post-hoc feature analysis [6]. A stratified random sampling-based method for interactive visualization of cosmology particle data was proposed by Woodring et al. [24]. For unstructured mesh, a zero copy in situ data structure was introduced in [25]. In another work, Ahrens et al. used an in situ image-based approach called Cinema [1] for flexible image-based post-hoc feature analysis. Instead of sampling data points for summarization, Cinema samples rendered data images across various visualization dimensions so that an image database can be generated for effective visual exploration. In situ distribution-based data reduction techniques were developed by Dutta et al. [8, 9]. An in situ analysis-driven data partitioning and representative sample selection was done in [17]. Data sampling based on bitmap indexing was proposed by Su et al. in [20]. Recently, Wei et al. [22] proposed an in situ data sampling approach which extended traditional stratified random sampling using bitmap indexing-based compressed data representation. For a more comprehensive view of the in situ data analysis techniques, please refer to the state-of-the-art report [3].

While sampling data points for producing visualization, Park et al. [18] proposed a technique which samples only a very small fraction of data for producing an accurate visualization. In the context of scatter and map plots, Park et al. designed a loss function which maximizes the visual fidelity of the scatter and map plots. However, the scope of this work is only limited to two such specific types of plots and the loss function proposed cannot guarantee high-quality samplings for producing general purpose visualizations. In another work, to improve upon random sampling for visualization, Nguyen and Song [16] proposed a centrality clustering based data sampling scheme. In contrast, we aim to propose a more generic in situ sampling scheme which can be used to generate a down-sampled data set that can produce visual representations of the data where the important features are preserved. Several researchers in the past have used information theory driven approaches for performing sampling of data sets. The primary goal of these bodies of works [5, 12, 19] was to select a subset of data which is most informative about the complete data set through maximizing the entropy of the selected subset. In this work, following the similar principle of information theory, we propose a scheme of sampling for selecting a set of informative data subset which preserves the important features in the data.

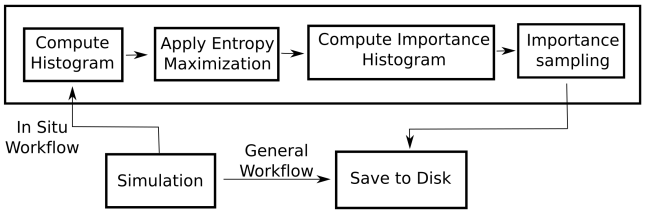
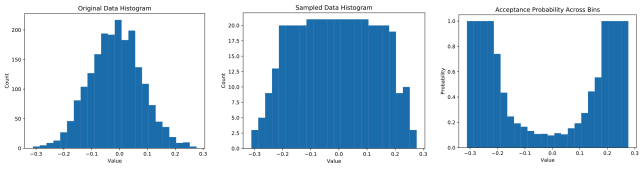


Figure 1: Schematic depiction of the proposed system for information-driven adaptive in situ sampling of the large simulations.



(a) Original data histogram (b) After entropy maximization (c) Acceptance histogram

Figure 2: Illustration of our proposed information-driven sampling method via entropy maximization. a) histogram of original data points, b) histogram of the sampled data where the entropy of resulting histogram is maximized c) Acceptance histogram showing the probability of the values of each bin getting accepted after sampling

3 IN SITU SAMPLING METHODS

3.1 Overview

For handling very large simulations, as opposed to the existing post-processing workflows that store the complete dataset, we propose in situ data-adaptive sampling method for effective data reduction. A schematic diagram of the proposed workflow is shown in Figure 1 where the in situ workflow depicts the information-driven sampling that maximizes the information gained from the sampled data and also preserves features of the datasets. For comparison purposes, we have also implemented the popular stratified random sampling [24], and regular sampling. We have integrated these methods into an in situ framework for the Nyx [2] cosmology simulation. Nyx is a massively-parallel, compressible hydrodynamics simulation framework that simulates the cosmos at large scales. Below we provide the details of our proposed information-driven sampling method and discuss about how it can be applied in situ.

3.2 Information-Driven Sampling

The aforementioned approaches (Stratified random and Regular sampling) assume that all samples are equally important and the resulting distribution of the sampled data generally remain similar to that of the original data distribution. However, generally for most of the scientific datasets, not all the data values are equally important to the domain scientists. Therefore, based on the notion that the rare data values are generally more important (foreground) than the frequently occurring data values (likely to be background), our proposed sampling scheme derives its core idea from the well-accepted field of information theory. In information

theory, entropy H is a measure of information content [7] of a random variable X and it is defined as: $H(X) = -\sum_{x \in X} P(x) \log_2 P(x)$, where $P(x)$ is the probability of occurrence of x, with $x \in X$. The principle of maximum entropy is built on top of the information content of a random variable and it states that the maximum entropy state of a random variable is the best representation when no other information is available [11]. As a generic formulation of an information-driven sampling technique, we apply this idea in situ. Assuming the normalized histogram of a data variable represents its probability distribution function, we formulate a sampling strategy that maximizes the data entropy after it is sampled along with the assumption that the rare data values to be more important (i.e., the data values with low occurrence probability will have a higher chance of getting selected). Since a uniform distribution has the highest entropy, our proposed sampling algorithm tries to accept data values in such a way that they are well distributed across the histogram bins.

An example of this is shown in Figure 2. In this case, we have taken 2000 randomly generated samples from a Gaussian distribution (mean=0.0, standard deviation = 0.1) and the corresponding histogram is shown in Figure 2(a). This data has an entropy of 3.9 bits (using 24 bins). Now, if we are going to keep 20% of the original data, then use of our entropy maximization-based sampling scheme will result in a sampled dataset with a histogram as shown in Figure 2(b). This sampled data has an entropy of 4.45 bits and as can be proven, for the given input histogram of original data [Figure 2(a)] and given the X samples to be drawn without replacement from that data, no other histogram is possible that has higher entropy.

Now, if we define an *acceptance histogram* as a histogram where the x-axis represents the data values and the y-axis is the probability of getting selected for the values falling into that bin,[e.g., Figure 2(c)] then it reveals that for entropy maximization, the data values which are rare in the original dataset (i.e., low count in the original histogram), will be given higher priority in the sampling process. Conversely, the frequently occurring data values (which often represent background or less important regions in the dataset), will have a much lower probability of acceptance. For our application, this generic method works quite well in preserving the important regions (features) of the data compared to the other two existing popular methods discussed above.

Starting from a regular grid dataset, this sampling method produces a particle dataset where for each particle, its location and density value is stored (similar to stratified sampling). On the contrary, the regular sampling method only needs to store the values since it produces regular grid data after sampling; i.e., given the same storage constraint, regular grid data can save 3× more data, but as we show below and in the next section, our proposed method still out-performs the regular and stratified random sampling method in preserving the features.

An example of the proposed method is shown in Figure 3 where the Hurricane Isabel dataset (for data description, visit: http://sciviscontest-staging.ieeevis.org/2004/data.html) is used for demonstration purposes. The most important feature of this dataset is the location and shape of the hurricane eye (e.g., Figure 3(a)) in the Pressure field. Using the Pressure field at time step 25, we applied the stratified random and our sampling schemes with a sampling ratio of 0.5% and regular sampling at 1.5% (since regular sampling

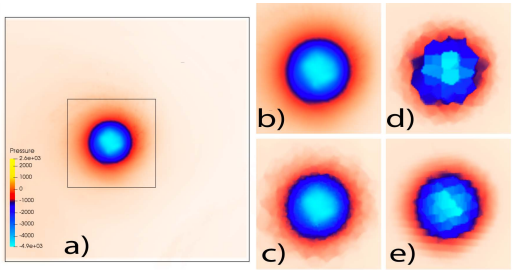


Figure 3: Sampling results from Isabel Hurricane dataset. a) original data b) zoomed in view of the feature (hurricane eye) region. c) reconstruction using our sampling method (sampling ratio 0.5%). d) reconstruction using random sampling method (sampling ratio 0.5%). e) reconstruction using regular sampling method (sampling ratio 1.5%).

Input: Sim generated data at each time step
Output: Acceptance histogram
n_k = samples to be taken from full data
N = total points from full data
nbins = number of bins
$n_{samps} = n_k \div nbins$: samples to be taken from each bin
remaining-samples = N
H = CreateHistogram(Data,nbins)
count,bin-edges = SortBinsAccordingToTheirCount(H)
while *not all bins are visited* **do**
  i=0;
  if *count[i]* < n_{samps} **then**
    samples-taken = count[i];
  else
    samples-taken = n_{samps};
  end
  P_i = samples-taken ÷ count[i]
  remaining-samples = remaining-samples - samples-taken
  remaining-bins = nbins - i
  n_{samps} = remaining-samples ÷ remaining-bins
  i=i+1;
end
Use P_is as the probabilities for the corresponding bins of the acceptance histogram

Algorithm 1: Creation of Acceptance Histogram In Situ

does not need to store the locations of samples). For comparison with the original data, we have reconstructed 3D volumes from the samples and applied the same color-map and transfer function for visualization. As can be clearly observed, given the original data in Figure 3, our proposed method (in Figure 3(c)) produces the closest representation compared to stratified random sampling (Figure 3(d)) and regular sampling (Figure 3(e)) given the same storage constraint.

3.3 Algorithm Details and Scalable Implementation

For in situ implementation of our proposed sampling method, the first stage consists of the computation of a data histogram. Given the

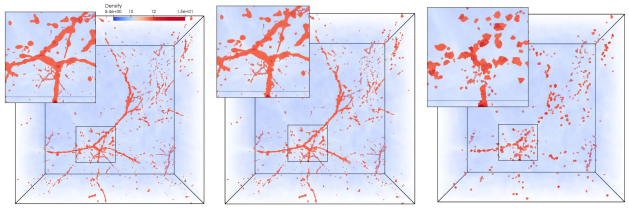


Figure 4: Volume visualization of Sampling results (sampling ratio 1%) from Nyx simulation. a) Original data and zoomed into one feature region. b) Reconstruction using our sampling method. c) Reconstruction using stratified random sampling method.

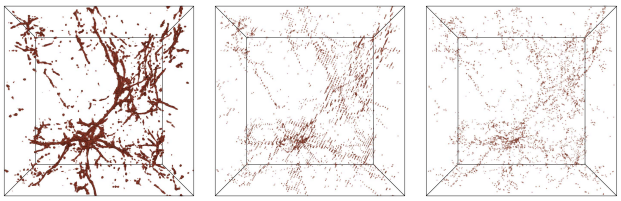


Figure 5: Point rendering results from Nyx simulation. a) using our sampling method (sampling ratio 0.5%). b) using regular sampling method (sampling ratio 1.5%). c) using stratified random sampling method (sampling ratio 0.5%).

in situ scenario, the data is generally distributed across processors and construction of global data histogram is performed using MPI-based communication across processors. This is scalable since each processor will compute its local histogram using the global bounds and then these local histogram will be added up in a MPI_Reduce method.

The next stage is the construction of an acceptance histogram based on this global histogram. To maximize entropy, the goal is to generate a near-uniform distribution from the samples. To achieve that, we first represent the histogram as a collection of tuples of count and bin ids. Next, we sort the histogram based on their counts in an ascending order. Since we know the total number of samples (n_k) we need to select given a sampling rate and also the total number of bins ($nbins$), the ideal uniform distribution would have n_{samps} samples where $n_{samps} = n_k/nbins$. If every bin in the original histogram had count value more than n_{samps}, then our target histogram would have $nbins$ bins with counts n_{samps}. This would maximize the entropy by generating a uniform distribution. Since not all bins of the original histogram may have count value more than n_{samps}, the target histogram will need to accommodate for this. The count of the i-th bin of the target histogram can only be $minimum(n_{samps}, count[i])$. As we are processing each bin starting from the bin with lowest count, the remaining samples will need to be accumulated and re-distributed across the remaining bins. A details of this one-pass algorithm is provided in Algorithm 1.

From this target histogram, acceptance histogram can be generated by computing the fraction of samples needed to be accepted for each bin by dividing the target histogram counts by the original histogram counts for each bin. The final stage of this process requires a second pass over the data points and accepting/rejecting them using that acceptance histogram. For each data point, now we first check which acceptance bin it falls and what is the acceptance probability. Using a random number generator, this data point can be accepted/rejected. This stage is also scalable once the acceptance histogram is generated and distributed across the processors.

4 COMPARISON AND PERFORMANCE

4.1 In situ Integration and Experimental Setup

To evaluate the in situ scaling performance of our proposed sampling method along with the other sampling methods, we conducted an in situ performance study using the Nyx cosmology simulation code [2]. Different sampling algorithms were implemented in C++ and were integrated into Nyx simulation code in the `writePlotFile();` in situ I/O routine. In situ sampling was performed during each simulation time step. The in situ performance study was conducted in a cluster with Intel Broadwell E5_2695_v4 CPUs (18 cores per node and 2 threads per core), and 125 GB of memory per node. Nyx cosmology simulation can generate data at various spatial resolutions depending on the input parameters. In our study, we ran Nyx which produced data of spatial resolution $512 \times 512 \times 512$ per time step.

4.2 Quality Comparison

The Nyx code generates a regular grid dataset where the density field is one of the important scalar fields and high-density regions generally indicate the Halos (features of interest). We provide visualizations of the in situ generated output data samples of different sampling schemes for comparison purposes in Figure 4 (using reconstructed volume visualization) and 5 (using direct particle rendering). From Figure 4, it is observed that when only 1% samples are kept, the proposed algorithm (Figure 4(b)) preserves the important features of the data quite well compared to the stratified random sampling (Figure 4(c)) where the original data is given in Figure 4(a). Given the same disk storage constraint, since regular sampling can store 3× more samples, we have compared our results with 3% samples of regular sampling. It is found that the regular sampling scheme still cannot preserve the features well, and similar observations can be applied to stratified random sampling. Figure 5 shows results for even lower sampling ratio 0.5% for our method in Figure 5(a), 1.5% for regular in Figure 5(b), and 0.5% for stratified random in Figure 5(c)). In the context of feature preservation, the proposed method outperforms the other methods even at such low sampling ratios.

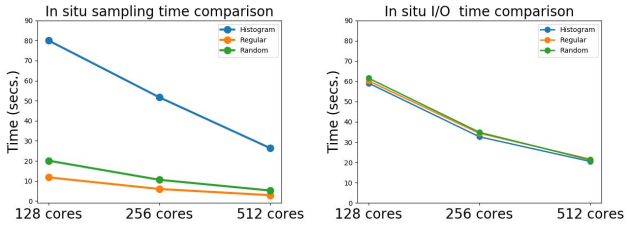
Next, to quantitatively compare the image quality produced by different sampling schemes to that of the original data, we use the Pearson's correlation coefficient as a comparison metric. We use the red, green, and blue channels of the images from each sampling method and compute the correlation coefficient with the image generated from the original data. The rendering parameters are fixed for all the cases. The result is presented in Table 1. As observed, the proposed method produces superior visual quality compared to the other two methods.

Table 1: Quantitative similarity comparison of different sampling method produced images to the image produced by the original data under same rendering configuration.

	Hist. Samp	StRand. Samp	Regular Samp
Isabel data	0.978	0.921	0.961
Nyx data	0.987	0.865	0.881

Table 2: In situ percentage timings of different sampling methods and I/O timings w.r.t the simulation timings.

	Hist. Samp (%sim time)	Reg. Samp (%sim time)	StRand. Samp (%sim time)	Hist. I/O (%sim I/O)	Reg. I/O (%sim I/O)	StRand. I/O (%sim I/O)
128 Cores	1.39	0.20	0.35	2.86	2.92	2.97
256 Cores	1.83	0.21	0.37	3.07	3.23	3.28
512 Cores	1.41	0.16	0.28	2.61	2.75	2.70



(a) (b)

Figure 6: a) Comparison of in situ sampling times among the proposed method (blue), regular sampling (orange), and random sampling (green). b) Comparison of in situ sampled data I/O times among the proposed method (blue), regular sampling (orange), and stratified random sampling (green).

4.3 In Situ Timing Comparison

To study the performance of our proposed sampling scheme, we conducted an in situ performance study using the Nyx simulation by running 100 time steps. Figure 6 shows the in situ sampling and I/O times for all three methods. As can be seen that with an increased number of cores, our method scales well along with the other two methods. It is also observed that compared to regular sampling and stratified random sampling, our method takes slightly longer time. However, given the quality of improvement we get in terms of preserving the important features in the sampled data, we believe the slight extra time is well justified. Further, the in situ I/O times in (Figure 6(b)) are similar for all the three methods. Next, in Table 2 we show the % of simulation time spent in performing the in situ sampling algorithms. As expected, the regular and stratified random sampling methods are faster compared to the proposed sampling method, but still, our method takes only around 1.5% of the simulation time on average which is negligible in the current context. Table 2 also reports the percentage disk I/O time for the sampled data with respect to the simulation raw data I/O and as can be seen that the I/O for sampled data is a small fraction of the simulation raw I/O.

5 DISCUSSION AND LIMITATIONS

The proposed sampling scheme currently works on each variable independently while sampling. However, scientific simulations often produce multiple variables and while sampling, the relationships among different variables need to be considered so that the variable relationships are preserved in the down-sampled data sets. Therefore, in the future, we would like to extend this sampling scheme into the multivariate domain. It is to be noted that by analyzing a multivariate distribution the similar ideas can be extended to the multivariate domain, however, high-dimensional distributions suffer from high computational cost and high storage requirements. Therefore, we will exploit compact distribution modeling schemes [4, 15] to reduce the analysis overhead in the in situ environment. Also, for conducting the evaluation and comparison studies, we have used the Nyx cosmology simulation which produced data of spatial resolution $512 \times 512 \times 512$. In the future, we plan to perform scaling study using other higher resolution data sets.

While evaluating the quality of the sampling schemes, we conducted an image-based comparison study. The motivation was to explore the quality of visualization that can be produced by different methods under a fixed rendering configuration. However, we would like to perform more detailed comparison study by directly comparing the 3D reconstructed data to the raw data. Finally, we acknowledge that the assumption that the rare values are more important is true for outliers and noise values as well. Hence, the proposed method will store such data values in the sampled data set. However, it is to be noted that, such outliers can be of importance to the scientists and by keeping such values we ensure that our sampling scheme preserves the diverse nature of the data.

6 CONCLUSION AND FUTURE WORK

In this work, we have presented a novel information-driven sampling technique for in situ applications. Our proposed method uses the data histogram for computing an importance function for each data point through the use of entropy maximization. Compared to several popular sampling methods, our proposed method preserves the features of the dataset much better, given a fixed dist storage budget. We further demonstrate the superior qualities of our in situ sampling method for Nyx cosmology simulation. In the future, we would like to explore the in situ methods for information-driven sampling opportunities for multivariate and multi-resolution datasets. We would further like to add data specific constraints for the proposed generic sampling method for more improved data handling and simulation steering.

ACKNOWLEDGMENTS

The authors wish to thank the anonymous reviewers for their insightful and detailed comments. This research was supported by the Exascale Computing Project (ECP), Project Number: 17-SC-20-SC, a collaborative effort of two DOE organizations - the Office of Science and the National Nuclear Security Administration.

REFERENCES

[1] J. Ahrens, S. Jourdain, P. OLeary, J. Patchett, D. H. Rogers, and M. Petersen. 2014. An Image-Based Approach to Extreme Scale in Situ Visualization and Analysis. In *SC14: International Conference for High Performance Computing, Networking, Storage and Analysis*. 424–434. https://doi.org/10.1109/SC.2014.40

[2] A. S. Almgren, J. B. Bell, M. J. Lijewski, Z. Lukić, and E. Van Andel. 2013. Nyx: A Massively Parallel AMR Code for Computational Cosmology. *apj* 765, Article 39 (March 2013), 39 pages. https://doi.org/10.1088/0004-637X/765/1/39 arXiv:astro-ph.IM/1301.4498

[3] Andrew C. Bauer, Hasan Abbasi, James Ahrens, Hank Childs, Berk Geveci, Scott Klasky, Kenneth Moreland, Patrick O'Leary, Venkatram Vishwanath, Brad Whitlock, and E. W. Bethel. 2016. In Situ Methods, Infrastructures, and Applications on High Performance Computing Platforms. *Computer Graphics Forum* (2016).

[4] J. Chanussot, A. Clement, B. Vigouroux, and J. Chabod. 2003. Lossless compact histogram representation for multi-component images: application to histogram equalization. In *IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No.03CH37477)*, Vol. 6. 3940–3942 vol.6. https://doi.org/10.1109/IGARSS.2003.1295321

[5] Xiang-Hui Chen, Arthur P. Dempster, and Jun S. Liu. 1994. Weighted Finite Population Sampling to Maximize Entropy. *Biometrika* 81, 3 (1994), 457–469. http://www.jstor.org/stable/2337119

[6] Hank Childs. 2015. Data Exploration at the Exascale. *Supercomputing frontiers and innovations* 2, 3 (2015). http://superfri.org/superfri/article/view/78

[7] Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience.

[8] S. Dutta, C. M. Chen, G. Heinlein, H. W. Shen, and J. P. Chen. 2017. In Situ Distribution Guided Analysis and Visualization of Transonic Jet Engine Simulations. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (Jan 2017), 811–820.

[9] S. Dutta, J. Woodring, H. W. Shen, J. P. Chen, and J. Ahrens. 2017. Homogeneity guided probabilistic data summaries for analysis and visualization of large-scale data sets. In *2017 IEEE Pacific Visualization Symposium (PacificVis)*. 111–120. https://doi.org/10.1109/PACIFICVIS.2017.8031585

[10] N. Fabian, K. Moreland, D. Thompson, A. C. Bauer, P. Marion, B. Gevecik, M. Rasquin, and K. E. Jansen. 2011pages = 89-96, doi = 10.1109/LDAV.2011.6092322,. The ParaView Coprocessing Library: A scalable, general purpose in situ visualization library. In *2011 IEEE Symposium on Large Data Analysis and Visualization (LDAV)*.

[11] E. T. Jaynes. 1957. Information Theory and Statistical Mechanics. *Phys. Rev.* 106, 4 (May 1957), 620–630. https://doi.org/10.1103/PhysRev.106.620

[12] Chun-Wa Ko, Jon Lee, and Maurice Queyranne. 1995. An Exact Algorithm for Maximum Entropy Sampling. *Operations Research* 43, 4 (1995), 684–691. https://doi.org/10.1287/opre.43.4.684

[13] Sriram Lakshminarasimhan, Neil Shah, Stephane Ethier, Scott Klasky, Rob Latham, Rob Ross, and Nagiza F. Samatova. 2011. Compressing the Incompressible with ISABELA: In-situ Reduction of Spatio-temporal Data. In *Euro-Par 2011 Parallel Processing*, Emmanuel Jeannot, Raymond Namyst, and Jean Roman (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 366–379.

[14] Jay F. Lofstead, Scott Klasky, Karsten Schwan, Norbert Podhorszki, and Chen Jin. 2008. Flexible IO and Integration for Scientific Codes Through the Adaptable IO System (ADIOS). In *Proceedings of the 6th International Workshop on Challenges of Large Applications in Distributed Environments (CLADE '08)*. ACM, 15–24. https://doi.org/10.1145/1383529.1383533

[15] K. Lu and H. Shen. 2015. A compact multivariate histogram representation for query-driven visualization. In *2015 IEEE 5th Symposium on Large Data Analysis and Visualization (LDAV)*. 49–56. https://doi.org/10.1109/LDAV.2015.7348071

[16] T. T. Nguyen and I. Song. 2016. Centrality clustering-based sampling for big data visualization. In *2016 International Joint Conference on Neural Networks (IJCNN)*. 1911–1917. https://doi.org/10.1109/IJCNN.2016.7727433

[17] B. Nouanesengsy, J. Woodring, J. Patchett, K. Myers, and J. Ahrens. 2014. ADR visualization: A generalized framework for ranking large-scale scientific data using Analysis-Driven Refinement. In *Large Data Analysis and Visualization (LDAV), 2014 IEEE 4th Symposium on*. 43–50. https://doi.org/10.1109/LDAV.2014.7013203

[18] Yongjoo Park, Michael J. Cafarella, and Barzan Mozafari. 2015. Visualization-Aware Sampling for Very Large Databases. *CoRR* abs/1510.03921 (2015). arXiv:1510.03921 http://arxiv.org/abs/1510.03921

[19] M. C. Shewry and H. P. Wynn. 1987. Maximum entropy sampling. *Journal of Applied Statistics* 14, 2 (1987), 165–170. https://doi.org/10.1080/02664768700000020

[20] Yu Su, Gagan Agrawal, Jonathan Woodring, Kary Myers, Joanne Wendelberger, and James Ahrens. 2013. Taming Massive Distributed Datasets: Data Sampling Using Bitmap Indices. In *Proceedings of the 22Nd International Symposium on High-performance Parallel and Distributed Computing (HPDC '13)*. ACM, New York, NY, USA, 13–24. https://doi.org/10.1145/2462902.2462906

[21] V. Vishwanath, M. Hereld, and M. E. Papka. 2011. Toward simulation-time data analysis and I/O acceleration on leadership-class systems. In *2011 IEEE Symposium on Large Data Analysis and Visualization (LDAV)*. 9–14. https://doi.org/10.1109/LDAV.2011.6092178

[22] T. Wei, S. Dutta, and H. Shen. 2018. Information Guided Data Sampling and Recovery Using Bitmap Indexing. In *2018 IEEE Pacific Visualization Symposium (PacificVis)*. 56–65. https://doi.org/10.1109/PacificVis.2018.00016

[23] Brad Whitlock, Jean M. Favre, and Jeremy S. Meredith. 2011. Parallel in Situ Coupling of Simulation with a Fully Featured Visualization System. In *Proceedings of the 11th Eurographics Conference on Parallel Graphics and Visualization (EGPGV '11)*. Eurographics Association, 101–109. https://doi.org/10.2312/EGPGV/EGPGV11/101-109

[24] J. Woodring, J. Ahrens, J. Figg, J. Wendelberger, S. Habib, and K. Heitmann. 2011. In-situ Sampling of a Large-scale Particle Simulation for Interactive Visualization and Analysis. In *Proceedings of the 13th Eurographics / IEEE - VGTC Conference on Visualization*. Eurographics Association, 1151–1160. https://doi.org/10.1111/j.1467-8659.2011.01964.x

[25] Jonathan Woodring, James Ahrens, Timothy J. Tautges, Tom Peterka, Venkatram Vishwanath, and Berk Geveci. 2013. On-demand Unstructured Mesh Translation for Reducing Memory Pressure During in Situ Analysis. In *Proceedings of the 8th International Workshop on Ultrascale Visualization*. ACM, Article 3, 8 pages. https://doi.org/10.1145/2535571.2535592