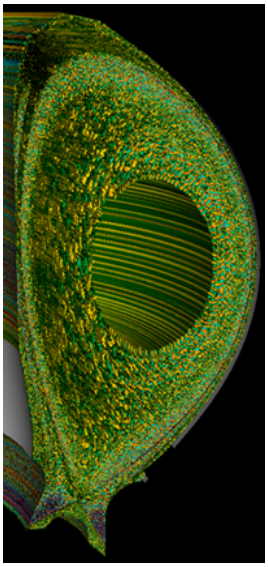
Data Reduction for Science: Brochure from the Advanced Scientific Computing Research Workshop

Scott Klasky, Oak Ridge National Laboratory
Jana Thayer, SLAC National Accelerator Laboratory
Habib Najm, Sandia National Laboratories

Publication date: April 15, 2021
Web DOI: 10.2171/1770192
DOE Office of Science Technical Contact: William Spotz (William.Spotz@science.doe.gov)

Introduction

The reduction of streaming and voluminous data sets while maintaining accurate representations of quantities of interest (QoIs) is a critical capability across the Office of Science (SC). SC-supported experiments, observations, and simulations produce data at volumes and velocities that are already overwhelming network, storage, and compute capabilities and their projected growth will greatly exacerbate this imbalance. The Advanced Scientific Computing Research program office held a virtual workshop in January 2021, bringing together 155 participants and 41 observers across experimental, observational, and computational application areas and research thrust areas in compression, reduced representations, experiment-specific triggers, filtering, and feature extraction/QoIs to identify priority research directions (PRD) leading to enhanced capabilities in data reduction. This workshop examined many scientific drivers, such as radio astronomy, fusion, combustion, climate, light sources, nuclear physics, and genomics, which are in desperate need for new Research & Development (R&D) in data reduction, because they currently risk ad hoc decisions that can limit the amount of knowledge gathered from SC facilities.



New workflows are beginning to emerge to both manage data and fully exploit the incredibly rich information produced by SC facilities. These data reduction workflows employ triggering, filtering, sampling, compression, reduced order modeling and feature detection. The workflows extend from observational/experimental devices to networks to remote and local storage to desktop and leadership computing facilities and require optimization across a diverse range of hardware.

In order for application scientists to trust data reduction methodologies, reduction techniques should be usable and adoptable by communities through best practices, benchmarks, data sharing, resource sharing, and through the development of tools that enable scientists to navigate these resources. The workshop focused on new R&D capabilities which can allow scientists to quantify the uncertainties in QoIs, along with preserving features to a specified tolerance. Furthermore, progressive techniques for streaming data need to be developed to enable scientists to make tradeoffs between the uncertainty, speed, and resource utilization. Since these workflows typically run on all types of

computational resources, these techniques need to be highly performant and portable across different architectures.

Science Drivers

Many science applications such as high energy physics, nuclear physics, radio astronomy, and light sources highlighted the large increase in data volumes. To satisfy network and storage constraints, experimental facilities reduce data at the edge before it is moved and stored. In some cases, in experimental and observational facilities, such as in high energy physics, nuclear physics, light sources, fusion, detectors and triggers are used to enhance the selection of scientifically interesting activity to save to storage and/or analyze during the experiment. For example, radio astronomy uses complex filters and triggers to determine which data are saved, a technique that not only allows scientists to store less data, but to also look for short-lived events, thereby linking reduction with anomaly detection. Simulations, such as in combustion and fusion can generate hundreds of petabytes of data in a day and need *in situ* reduction techniques.

New detector technologies allow more data to be captured at higher rates, which increases the data velocities, and necessitates the need for streaming data reduction techniques. Simulations use accelerators to increase the velocity of their calculations, which increases the velocity of the data being generated. For example, computational combustion relies on large simulations that search for features, events, and anomalies and produce QoIs that could be combined with in-situ machine learning and artificial intelligence workflows to produce reduced order models in addition to a complete data model repository. Next generation experimental and observational facilities are also generating data at increased velocities and often require fast feedback on data-taking, placing hard real-time requirements on data reduction. By providing scientific information within seconds, anomaly detection could take place in near real time, allowing material characterization experimenters to collect data under optimum conditions. In all of these cases, there is a crucial need for fast reduction techniques that must work on diverse architectures and stream data across processes in complex experimental workflows, ensuring that short-term events (for example from the Very Large Array and Square Kilometer Array radio astronomy facilities) can be captured and analyzed in the reduction process.

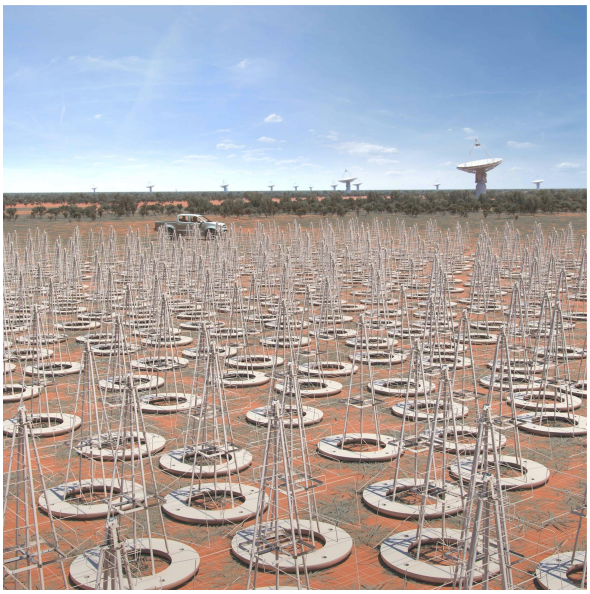
Data from these facilities are too expensive or too difficult to reproduce. Capturing the provenance is viewed as an essential part of the reduction process to foster trust and add value to the data. Ensuring that the saved reduced data can be trusted requires, among other things, quantification of uncertainty in the QoIs, a significant challenge given the large variety of data each science case produces. In all of the science cases, several design issues must be met. 1) The reduction techniques must be trustworthy so that all of the post-processing of the data contains no inaccurate results or non-physical artifacts. 2) Reduction methods need to be efficient and accurate, and scientists need to understand the tradeoff between efficiency and accuracy, and 3) The techniques need to work in science workflows, satisfying constraints specific to each workflow's deployment environment (e.g., "edge," HPC resources).

The workshop led to the formulation of the following four priority research directions (PRDs).

PRD1: Effective and trustworthy reduction algorithms and tools

A recurring theme among applications discussed at the workshop, such as fusion simulations and next-generation radio astronomy facilities, was the vital importance of understanding and quantifying the impact of data reduction on the principal outputs and QoIs produced by the application. The ability to quantify those impacts with realistic numerical bounds is essential if the scientist is to have confidence in applying data reduction. Ideally, the data reduction procedure should be able to preserve the accuracy in QoIs to any user-prescribed tolerance while, at the same time, not creating a bottleneck in the analysis.

Data reduction procedures should provide the scientist with the flexibility to specify certain user-prescribed quantities which should be preserved to a specified tolerance. Additionally, it is important to quantify the uncertainties in the QoIs and/or to specify the uncertainty before the data are reduced. This is essential to ensure that scientists can trust the reduction routines and alleviate concerns that the reduced data might fail to preserve physical invariants, such as mass, momentum, energy, etc., compromising the physical validity of the subsequent analysis or even creating non-physical features and instabilities in a simulation. There are many methodologies to reduce data, such as functional data analysis, statistical methods, hierarchical methods, and machine learning techniques. Although these techniques typically work well for uniform grids with quantifiable errors on the quantity being reduced, a key priority is ensuring that these can also work for nonuniform grids as well as being able to quantify the uncertainties in the QoIs.



While there is a clear need for reduction techniques that preserve accuracy, such techniques must be executed on machines subject to a limited set of resources such as memory, bandwidth, storage, and time. Accordingly, algorithms should be efficient in their use of resources and should permit clear understanding of the tradeoffs between achievable accuracy and resource constraints.

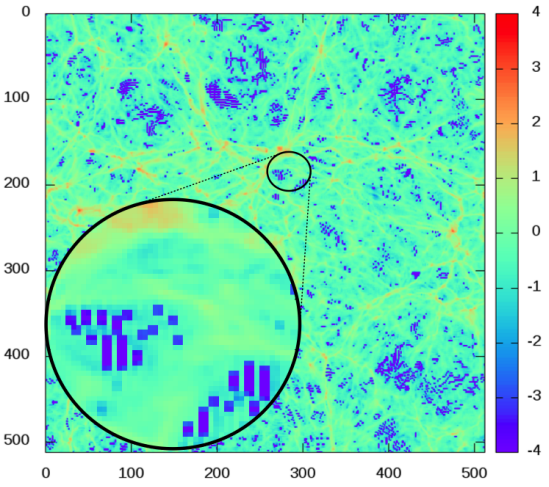
Research in this area should include the development of one or more of the following: 1) Effective algorithms with robustly estimable or provable error bounds for the data being reduced, 2) Effective algorithms with robustly estimable or provable error bounds in QoIs, 3) Algorithms which can meet constraints on time, memory, storage, and computational resources, 4) Understanding and optimizing tradeoffs between accuracy, memory usage, reduction error, and time to solution, 5) Methods for space-time, unstructured data (e.g., particles) and/or high-dimensional datasets, 6) Techniques to quantify uncertainty in the QoIs.

PRD2: Progressive data reduction to enable prioritization for efficient streaming

As data volumes and velocities increase, scientists need effective methods to move and process data progressively. Such techniques allow data to be refactored, according to scientific priorities, e.g. timeliness, freshness, accuracy, and uncertainty, so that they can be used for streaming data from

experiments or simulations, combining data from multiple sources, or retrieving data from storage. In addition, data reduction algorithms may take into account additional requirements from users based on the uncertainties of the input and output, as well as others including network bandwidth and computational constraints.

One key feature differentiating progressive data reduction from other reduction algorithms is that it is based upon user and system defined constraints. The most prominent constraint is time, whether for the data reduction step or the entire workflow. If scientists are willing to use extra time, then they may gain in accuracy or another measure. There are also situations in which there are constraints on the availability of computational resources (e.g. network bandwidth, free compute nodes, storage) during a live experiment/simulation. In these cases, users need to constrain the amount of data being moved and processed and only work with the most important data records based on their priorities. In other use cases, progressive reduction techniques may allow users to interactively select the amount of data that can be analyzed and visualized, allowing scientists to cope with system constraints, while providing assessment of uncertainties in their analytics.



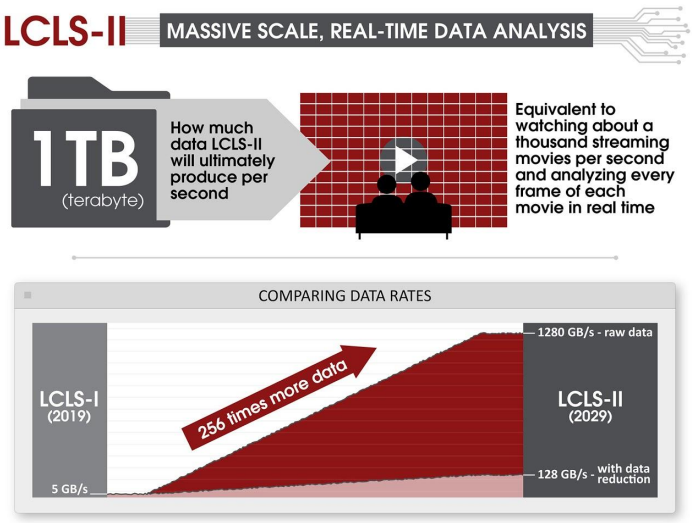
Research in this area should include the development of one or more of the following: 1) Developing effective reduction algorithms that allow for progressive data reduction and analysis taking into account the constraints mentioned above, 2) Understanding and developing uncertainty and reliability of progressive data reduction algorithms, 3) Developing adaptive algorithms that could dynamically adjust the accuracy based on the resource constraints, 4) Developing reliable execution time bounds for different error characteristics of downstream workflows, 5) Automating data reduction for streaming data taking into account the constraints and accuracy requirements of the end-to-end workflows.

PRD3: Algorithms that preserve information in features and quantities of interest

Research is needed to target the development of data reduction methods that preserve information and enable knowledge discovery directly from reduced representations. This includes development of scalable statistical methods for discovering/capturing underlying structure/manifolds in data, and that have data-reconstruction ability with theoretical bounds and/or quantified uncertainties. This is useful in order to enable performance guarantees for specific downstream QoIs and adaptation to user-specified tolerances. These methods should enable both interpolation and extrapolation of data, while preserving relevant structure, constraints, or physical invariants. Also necessary are methods that maximize preservation of high-level important information and minimize preservation of low-level noise and other artifacts (e.g., numerical discretization) unrelated to the QoIs. This can involve, for example, data transformations to enhance/separate useful or unanticipated information from less useful/known information. R&D should also target methods that enable deriving representations, features, and QoIs that can be used as inputs to subsequent downstream operations such as querying, subsetting, feature

detection, feature tracking, summarization, projection, and further transformations. Work is also needed to provide metrics that reflect application-specific information, which can be used in the development of algorithms for compression and data reduction.

Moreover, in order to provide measures of confidence in the quality of reduced data representations, work should target the broad landscape of error analysis and uncertainty quantification in data-reduction workflows. This includes work to estimate the bounds on error, as well as the form, distribution, and magnitude of noise and uncertainty, in order to assess the amount of information that can be reliably extracted from a given dataset, and captured in reduced representations. Further, there is a need for optimal reduction methods, targeting reduction based on sufficient statistics, that preserves goal-oriented information in specific features and/or QoIs.



Research in this area should include the development of one or more of the following: 1) Reduction methods that learn and preserve underlying structure and physical invariants in data. 2) Reduction methods that provide effective representation of high-level information on QoI while robustly handling noise and other artifacts. 3) Goal oriented methods targeting reduced representations with data-reconstruction ability, tailored for specific downstream operations. 4) Reduction methods that provide quantification of uncertainty in reduced representations, accounting for errors, noise, as well as data artifacts. 5) Methods for learning optimal reduced data representations, with parsimonious capture of maximal information.

PRD4: Mapping to new architectures and use cases

Scientific instruments producing vast amounts of data often use systems composed of tiered layers of hardware and software to enhance the selection of scientifically interesting activities for saving to persistent storage or to analyze during an experiment. The data may be acquired from multiple sources with varied readout timescales, and associated pipelines require real-time or near real-time reactions to keep up with the dataflow. With the next generation of electronics, networking, computing, and storage technologies, and advances in data reduction technologies, there is a possibility to rethink how these pipelines are architected, reflecting the need to operate in ever more extreme environments.

Novel technologies and emerging architectures provide new opportunities to address these data reduction requirements and also lead to new research challenges. As the complexity and heterogeneity of technologies such as purposeful accelerators, programmable edge and in-network computing, and new storage devices grow, new research is needed in data reduction algorithms and software stacks that can leverage their unique capabilities. Ensuring portability across these highly heterogeneous architectures will be increasingly important. These data reduction algorithms would need to meet hard performance

requirements on real-time latency, computation, network bandwidth, or compression factors, depending on the application.

Research in this area should include the development of one or more of the following: 1) Novel reduction techniques implemented on heterogeneous architectures capable of extracting information that distills data into actionable information on the timescales required by experiment, 2) Development of composable data reduction pipelines and workflows. 3) Operation and optimization of reduction and feature extraction techniques under varying real-time and near-real-time constraints, implemented in different layers of hardware, software, and executing on the edge, in the network, and on HPC resources. 4) Development of portable and adaptable reduction techniques that can enable the autonomous control of experiments. 5) Co-design of highly-optimal reduction algorithms on current and next generation hardware.

Summary

Data reduction for science holds promise for addressing the challenges of moving, storing, and processing massive data sets produced by the scientific community. Pursuing the PRDs outlined here will enable advances in data streaming, fast feedback and/or autonomous control of experiments, and faster time to scientific insight. These advances will result in a significant improvement in the ability to transport, store, process and interpret experimental, observational, and computational data.

Workshop Executive Committee

Scott Klasky, Oak Ridge National Laboratory; Jana Thayer, SLAC National Accelerator Laboratory; Habib Najm, Sandia National Laboratories

Workshop Organizing Committee

Mark Ainsworth, Brown University; Amber Boehnlein, Jefferson Lab; Stuart Campbell, Brookhaven National Laboratory; Kevin Carlberg, University of Washington; Hank Childs, University of Oregon; Ian Foster, Argonne National Laboratory; Jeff Kern, National Radio Astronomy Observatory; Youssef Marzouk, Massachusetts Institute of Technology; Manish Parashar, National Science Foundation; Heidi Sofia, National Institutes of Health; John Wu, Lawrence Berkeley National Laboratory