Computing and Visualization in Science manuscript No.
(will be inserted by the editor)

ND^{2}AV: N-Dimensional Data Analysis and Visualization

Analysis for the National Ignition Campaign

Peer-Timo Bremer · Dan Maljovec · Avishek Saha · Bei Wang · Jim Gaffney · Brian K. Spears · Valerio Pascucci

Received: date / Accepted: date

Abstract One of the biggest challenges in high-energy physics is to analyze a complex mix of experimental and simulation data to gain new insights into the underlying physics. Currently, this analysis relies primarily on the intuition of trained experts often using nothing more sophisticated than default scatter plots. Many advanced analysis techniques are not easily accessible to scientists and not flexible enough to explore the potentially interesting hypotheses in an intuitive manner. Furthermore, results from individual techniques are often difficult to integrate, leading to a confusing patchwork of analysis snippets too cumbersome for data exploration. This paper presents a case study on how a combination of techniques from statistics, machine learning, topology, and visualization can have a significant impact in the field of inertial confinement fusion. We present the ND^{2}AV: N-Dimensional Data Analysis and Visualization framework, a user-friendly tool aimed at exploiting the intuition and current workflow of the target users. The system integrates traditional analysis approaches such as dimension reduction and clustering with state-of-the-art techniques such as neighborhood graphs and topological analysis, and custom capabilities such as defining combined metrics on the fly. All components are linked into an interactive environment that enables an intuitive exploration of a wide variety of hypotheses while relating the results to concepts familiar to the users, such as scatter plots. ND^{2}AV uses a modular design providing easy extensibility and customization for different applications. ND^{2}AV is being actively used in the National Ignition Campaign and has already led to a number of unexpected discoveries.

Keywords topological analysis · visualization · dimension reduction

P-T. Bremer
Lawrence Livermore National Laboratory
E-mail: bremer5@llnl.gov

D. Maljovec
Scientific Computing and Imaging Institute, U. of Utah
E-mail: maljovec@cs.utah.edu

A. Saha
Yahoo Labs
E-mail: avishek@cs.utah.edu

B. Wang
Scientific Computing and Imaging Institute, U. of Utah
E-mail: beiwang@sci.utah.edu

J. Gaffney
Lawrence Livermore National Laboratory
E-mail: gaffney3@llnl.gov

B. Spears
Lawrence Livermore National Laboratory
E-mail: spears9@llnl.gov

V. Pascucci
Scientific Computing and Imaging Institute, U. of Utah
E-mail: pascucci@sci.utah.edu

1 Introduction

Some of the most exciting and challenging frontiers of science are the ultra large-scale experimental facilities such as the Large Hadron Collider at CERN or the National Ignition Facility (NIF) at Lawrence Livermore National Laboratory. These facilities allow us to explore physics in regimes far beyond previous capabilities and virtually by design exceed the limits of our theoretical understanding. To bridge the gap between our current knowledge and the observed outcomes, the experiments are typically accompanied by an equally impressive effort in developing predictive simulations of various fidelity. Such simulations are used to design the experiments and to plan the facilities, to re-create

observed phenomena to better investigate unobservable aspects of an experiment, and to explore, as a test bed, the influence different physical models may have on the predicted outcome. Nevertheless, it is quite common for even the most sophisticated simulations to deviate significantly from the corresponding experiments, a clear indicator of our still insufficient understanding of physics. Together, the experimental results and the simulation data form a complex mix of information at various scales and fidelity that is often accumulated over decades and represents investments of potentially billions of dollars.

To gain new insights and discover new physics from this data collection represents a challenge quite different from more traditional scientific analysis problems. When analyzing a medical scan, a climate simulation, or a simulated fluid flow, there typically exists a more or less specific question a scientist would like to explore. More importantly, the answer is expected to be contained within the simulated data. Instead, in complex systems such as NIF, the only known fact is that the best current models and theories are insufficient to explain reality. The corresponding gap may be due to faulty theories, inadequate numerical models, or yet unknown physical effects. As a result, the current analysis process relies heavily on the intuition of highly trained scientists and engineers to form and test new hypotheses in order to better explain the underlying phenomena. To gain the necessary insights, scientists will typically search for previously unknown or yet unexplained interactions between the inputs and outputs of the simulations, among the experimental data, or between both sets. However, currently this process is severely limited by our ability to detect such relationships, especially in the large, disparate, and high-dimensional datasets of greatest interest.

This paper presents a case study on how a visual-analysis-driven approach that integrates various high-dimensional analysis techniques with an interactive visual interface can lead to important new insights. In particular, we introduce ND^2 AV, an interactive environment to explore, analyze, and visualize high-dimensional and multivariate data. The framework combines a number of established analysis techniques, such as dimension reduction and clustering, with state-of-the-art techniques in topological analysis and high-dimensional neighborhood graphs. The former provide insights into the shape and structure of a domain of interest and the latter explore the structure of a particular quantity of interest with respect to this domain. ND^2 AV, for the first time, combines these two aspects of data analysis and allows scientists to construct a more complete picture of their data. More importantly, the system, while generally applicable, has been specifically designed to extend rather than replace the current scientific workflow and to integrate the intuition of the domain scientists as much as possible.

The system relies on simple and intuitive drag-and-drop techniques to allow even novice users to quickly create sophisticated hierarchical workflows. All modules in a workflow are automatically cross-linked, allowing users to intuitively explore the influence any parameter choice in one module has on any other result. In this manner, ND^{2}AV provides an interactive feedback that is vital in supporting a smooth and seamless exploration process. Most results presented here could have been achieved through a clever combination of existing tools, but only the ability to effortlessly browse the vast array of possibilities allowed scientists to find the relevant components. Finally, ND^2 AV is easily extendable to include further analysis tools and readily adaptable to a large number of application areas.

In the case study presented here, we discuss the role ND^{2}AV is playing in the National Ignition Campaign (NIC), its original target application. The NIC, a collaboration among Lawrence Livermore, Los Alamos, and Sandia National Laboratories as well as The University of Rochester and General Atomics, is aimed at demonstrating inertial confinement fusion (ICF), that is, thermonuclear ignition and energy gain in a laboratory setting. As will be described in more detail below, the goal is to focus 192 beams of the most energetic laser built so far onto a tiny capsule containing frozen deuterium. Under the right conditions, the resulting pressure will collapse the target to the point of ignition where hydrogen starts to fuse and produce massive amounts of energy, effectively creating a small star.

The NIC has made significant progress, but the ultimate goal – ignition – has not yet been reached. One of the primary challenges is that current simulation codes based on the latest physical models disagree in some fundamental aspects with the experimental results, making tuning the experiments extremely challenging. The discrepancies are widely believed to be the result of our yet limited understanding of matter and energy under extreme conditions (i.e., at the center of a star). However, an NIF shot is an exceedingly complex, highly coupled, and markedly nonlinear process and at this time it is unclear whether our current models are lacking, employ invalid assumptions, or are simply incorrect. As such, there is a considerable effort being spent in analyzing various simulation models, either of the entire system or of individual components, in par-

ticular as they relate to the experiments performed so far.

ND^2 AV has been designed to support this effort and has already led to several unexpected discoveries. In particular, a widely held assumption is that in the high-dimensional parameter space of possible outcomes there exists a single "plateau" of yield (the amount of energy produced in a shot). Much of the past effort has been focused on finding a way "uphill" to achieve ignition. However, our analysis of several simulation ensembles suggests that the response of the models is far more complex and localized than previously expected. As will be discussed in Section 5, many quantities of interest show decidedly local behavior with different correlations among variables in different portions of the parameter space. The ability to quickly browse such data provided by ND^{2}AV has significantly decreased the amount of time and effort necessary to discover these cases of unexpected behavior.

According to (Munzner, 2009), any visualization design includes four levels: characterization of the problem domain, design of the data and operation abstraction, design of the visual encoding of the data and the interaction with the user, and finally algorithm design. The algorithms and visualization techniques used have been established in prior works; thus our contributions are in the first two categories: domain problem characterization and the specific operations we have decided to employ for analyzing the data. Therefore, we make a great effort to describe the challenges of bridging the gap between high-end physicists and the data analysis and visualization community. This work presents a first step in the right direction by presenting the ND^2 AV framework. The clustering, dimension reduction, geometric analysis, and topological analysis performed are standard, but the ease of use with which we combine them and allow for other techniques to be used in addition is what we believe to be a novel contribution at the second layer of the visualization design system.

In the remainder of the paper, we will first discuss related work in visualizing high-dimensional data, and then describe the NIC and in particular the different types of simulations used in our case study in more detail. We will next introduce the tool and its underlying design philosophy, followed by an in-depth description of several use cases and the results produced.

2 Related Work

We first review software systems that guide users through multivariate and high-dimensional data exploration and analysis, by encoding a wealth of techniques in an interactive visual environment. XmdvTool (Ward, 1994) offers visual exploration of multivariate data, by integrating multiple analysis and visualization techniques that focus on n-dimensional projection, such as standard graphical presentations (e.g., scatter plot matrices, glyphs, parallel coordinates), hierarchical clustering of dimensions, brushing, and linking. Orca (Sutherland et al, 2000) is a extensible toolkit designed for constructing interactive and dynamic linked data viewers for rendering, manipulating, and linking. GGobi (Cook and Swayne, 2007) (with plugin designed for R (R Development Core Team, 2008)) focuses on providing expert users (e.g., statisticians and scientists) interactive, high-dimensional data investigation tools, such as a set of plot types including scatter plot matrices, projection pursuit, and grand tour, and a set of manipulations such as brushing, scaling, visual and algorithmic (e.g., hierarchical) clustering, supervised and unsupervised classification, and statistical inference. Based on similar concepts, Mondrian (Theus and Urbanek, 2008) is another statistical data-visualization system, which in particular works well with categorical, geographical, and large data, providing advanced data selection/query techniques (e.g., boolean functions, selection sequence manipulations, selection rectangles), plots (e.g., mosaic plots, missing value plot), and linked analysis.

The prototype system developed by Guo (Guo, 2003) describes a human-centered exploratory environment that combines a suite of coordinated visualization and analysis components, centered around identifying interesting subspace via interactive feature selection and searching arbitrary-shaped multivariate clusters via hierarchical clustering. The rank-by-feature framework (Seo and Shneiderman, 2005) guides users to visually inspect and find important features using certain ranking criteria with axis-parallel 1D and 2D projections of multidimensional datasets (e.g., hierarchical-clustering-based or scatter-plot-based ordering). The system introduced in (Johansson and Johansson, 2009) combines user-defined quality metrics to preserve important features during dimension reduction, and offers automatic ordering of variables to enhance perception of patterns selected by the user. VisuMap (VisuMap Technologies Inc., 2009) is designed for visual exploratory analysis of a high-dimensional dataset, which includes mapping and dimension reduction (e.g., Multidimensional Scaling, Relational Perspective Map (Li, 2004)), clustering (e.g., k-means, affinity propagation), linked data views, scripting, and library interfaces for advanced applications. (Tatu et al, 2009) presents relevance measures (based on correlation and cluster separation) for typical analysis tasks based on scatter plots and

parallel coordinates, to assist the user in potentially finding relevant visual structures and speeding up the exploration process. For a survey on such quality metrics in guiding high-dimensional data visualization, see (Bertini et al, 2011).

DimStiller (Ingram et al, 2010) is a system that focuses on dimensionality reduction and analysis by providing local and global guidance to nonexpert users, through expression and operator abstractions that encapsulate a sequence of transformations acting upon tables of data, workflows that bundle together commonly used patterns of analysis, and immediate visual feedbacks via linked views and control panels guiding intrinsic dimension estimation and data exploration. In terms of topological methods, persistence-based clustering (Chazal et al, 2011) and Mapper (Singh et al, 2007) have been proposed to construct useful combinatorial representations for the analysis and visualization of high-dimensional datasets. Such techniques could be potentially integrated into a guided, interactive visual environment.

The goal of the aforementioned techniques is to explore and visualize multidimensional scalar functions. However, our proposed tool aims to visualize a function as well as validate and compare its performance with respect to models built using data collected. In the remainder of this section, we discuss a body of work that lies closer to our approach of using one or more data representations with an end goal of reconciling simulation results with predicted outcomes.

Hyperslice (van Wijk and van Liere, 1993), one of the earlier works in the realm of scientific visualization and interaction, uses a novel representation of multidimensional functions as orthogonal 2D slices that leads to faster rendering and ease of visual representation. Hypermoval (Piringer et al, 2010) was proposed to validate regression models (for car engine design) using a combined visualization of high-dimensional functions and available validation data. The tool World Lines (Waser et al, 2010) builds on the concept of computational steering and allows users to interactively add new information during the analysis process while combining multiple simulations via linked views to choose a final outcome. Vismon (Booshehrian et al, 2012) allows interactive visualization of multidimensional relationships between input and output variables targeted specifically for fishery applications. Built along similar lines, both Paraglide (Bergner et al, 2013) and Tuner (Torsney-Weir et al, 2011) allow the added advantage of iterative parameter tuning of the underlying model via interactive visualization tools.

For complex simulations such as diesel common injection systems, (Matkovic et al, 2005) proposes to use visualization tools to augment numerical optimization methods for analysis of simulation data and for investigating the effect of model parameter changes. (Berger et al, 2011) performs visualization-based interactive exploration of the parameter space guided by uncertainty in prediction of the underlying statistical model.

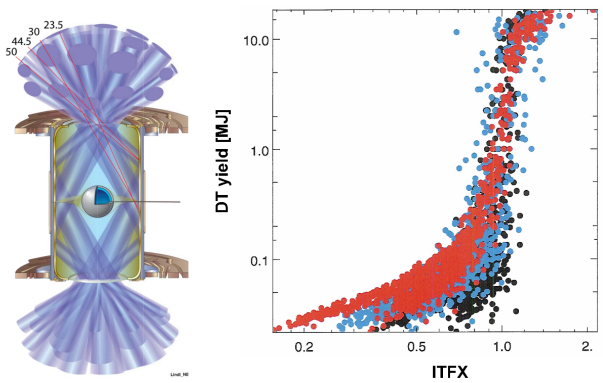
Compared to these related works, ND^2 AV not only includes standard graphical presentations (e.g., scatter plot views, parallel coordinates), traditional dimensionality reduction, and clustering algorithms, but also, for the first time, incorporates topological analysis techniques, e.g., (Correa et al, 2011; Gerber et al, 2010). As noted, for example, by (Correa et al, 2011), a topology-based segmentation provides a novel view into the structure of high-dimensional functions using intuitive notions such as segments formed around maxima or minima equivalent to high-dimensional "mountains" and "valleys." Techniques proposed by (Gerber et al, 2010) apply regression to summarize the topological structure of each cell of the approximated Morse-Smale complex (Edelsbrunner et al, 2003) of discretely sampled high-dimensional scalar fields. In particular, the direct interplay between these novel techniques and existing approaches has proven extremely fruitful and is a unique feature of ND^2 AV. Finally, our system offers not only data exploration guidance for nonexperts, but also a highly extensible module environment for intermediate level users with some programming background.

3 National Ignition Campaign (NIC)

Following the invention of the laser in 1960, physicists have been postulating ways to produce nuclear fusion using laser light as the primary energy driver (Kidder, 1974). After a decades-long pursuit, nuclear fusion ignition is within reach for the first time at the National Ignition Facility (NIF) at Lawrence Livermore National Laboratory (LLNL). NIF is the world's largest and most energetic laser, capable of delivering up to 2 MJ of laser energy to the target. Since 2009, NIF has been firing 192 laser beams occupying a football-stadium-sized facility to simultaneously illuminate a millimeter-scale fusion target. The objective of NIF experiments is to navigate a large space of engineering parameters, and to find the region of parameter space predicted by simulation that leads to successful inertial confinement fusion (ICF) (Lindl, 1998). The experimental ICF effort at the NIF represents one of the largest scientific endeavors being undertaken in the world today.

3.1 Inertial Confinement Fusion

ICF is an attempt to compress nuclear fuel, usually the hydrogen isotopes deuterium and tritium (DT), to pressures and temperatures high enough to force the hydrogen species to fuse. This fusion reaction yields high-energy neutrons and charged helium nuclei (alpha particles) that can be harnessed for energy production. The ICF fusion scheme uses the fuel's own inertia to prevent it from disassembling, thus providing the confinement necessary for the fuel to burn. ICF attempts at NIF use laser light to heat the interior of a small gold cylinder, or hohlraum. The hohlraum absorbs the incident laser and re-emits X-ray radiation that violently heats the outside of a spherical capsule located at the hohlraum center (see Fig. 1(a)). The exterior of the capsule is rapidly vaporized in a process known as ablation. The gas, or plasma, rapidly blows away from the surface, producing a rocket-like reaction that implodes the capsule. While imploding, the cryogenic DT ice layer inside the capsule accelerates to velocities approaching 350 km/s. When the imploding shell stagnates on itself near the capsule center, pressures in the hottest material reach 300 Gbar, leading to fusion burn. The entire assembly at this point is less than 100 um in diameter, is smaller than a human hair, and has reached instantaneous temperatures in excess of 50 million degrees. The charged alpha particles that are born in the hot spot self-heat the cooler, dense nuclear fuel, triggering a burn wave that propagates through the DT fuel until the assembly loses confinement and explodes.



(a) (b)

Fig. 1 (a) Schematic of an ignition target, the incoming 192 laser beams, and the surrounding hohlraum. (b) Yield can be ordered by various performance metrics. Here, ITFX is shown in red, ITF in blue, and the central hot spot pressure in black. All three common metrics show the abrupt transition from low to high yield at the cliff where the dimensionless metrics are defined to be 1. The methods developed in this paper reveal a more complicated structure that is blurred away by these physically motivated metrics.

The pursuit of ICF ignition depends heavily on detailed numerical simulations of implosion experiments using radiation-hydrodynamics codes. These types of simulations have proven to be predictive of the results at laser facilities smaller than NIF. While the scales of NIF experiments represent a challenge for modern computational efforts, one strategy adopted by the NIF target physics program has been to produce large ensembles of implosion simulations. The essential goal of these ensembles is to map out the capsule performance over the region of parameter space where experiments are expected to occur. Then, an essentially iterative approach is employed: perform experiments, use post-shot simulations to understand the experimental results in the context of the numerical database, and field new experiments with parameter settings that are expected to improve performance. In this way, the NIF ICF program attempts to "tune" the experiments until success is achieved.

This tuning notion requires the clear elucidation of a set of input or engineering parameters that can be adjusted to affect performance. It also requires a well-defined set of measurements or observations of the experiments that will provide feedback for comparison with simulation. For a system as complicated as NIF, any number of parameters can be defined as inputs – those affecting target dimensions and materials, those affecting initial densities and pressures, those affecting laser performance, and so on. We will focus here on parameters that affect the time-history of the laser brightness as it is delivered to the target. This tailoring of the laser intensity in time is known as pulse shaping, and it is used to set up a sequence of shock waves that precondition, or stiffen, the DT fuel before accelerating it for implosion (Robey et al, 2012). The pulse is shaped such that a sequence of four spherical shocks merges at a carefully chosen radial location in the target. Mis-timing of these shock collisions greatly reduces the compressibility of the DT fuel and represents a well-explored failure mechanism for the implosion.

The performance of the experiment is judged using a set of exquisitely precise diagnostic instruments consisting of neutron spectrometers, high speed X-ray microscope cameras, and various other nuclear and X-ray tools. A fundamental performance metric is the total energy produced by fusion reactions. Because this energy is associated with the production of neutrons, the yield of the reactions can be measured by counting 14 MeV neutrons with spectrometers. Experiments on the NIF are currently capable of producing just under 1e15 neutrons. To demonstrate self-heating by alpha particle deposition, this neutron yield must be seen to be just over 1e16. Upon ignition, a NIF target will pro-

duce at least 3-4e17 neutrons. It can be said that NIF experiments are currently searching for the yield cliff in order to drive it up. The rise in yield as performance improves is strongly nonlinear, and is often described as a cliff (Fig. 1(b)). Consequently, the neutron yield may be amplified greatly from a small improvement in performance, taking the yield from the foot of the cliff to the summit.

Yield is, of course, a multivariate function with many independent variables. The standard picture of the yield function presumes that there is a single global maximum surrounded by cliff behavior as the viewer moves away from the maximum in any direction. This smooth view of the performance space is supported by coarse, but reliable, physics principles that guide the understanding of both experimentalists and numerical theoreticians. Testing this notion of smooth topography requires a very large number of simulations to explore the high-dimensional parameter space. Importantly for the discussion here, it also requires tools and techniques to analyze and visualize the behavior of a function such as yield in the high-dimensional parameter space. It must also be emphasized that the results must be communicated clearly among a large group of skilled scientists with broad and diverse backgrounds.

3.2 Engineering/Macro-Simulations

Fundamentally, the goal of NIF is to probe the parameter space to find the region that leads to near-optimal performance. To perform this search empirically, one would require more experiments than could ever be performed at the NIF. Since 2009, NIF has shot just under 40 cryogenic layered implosions (Lindl et al, 2011), an outstanding achievement by the NIF team, which, however, will not allow for the exploration of sizable parameter spaces. The NIF program then relies on ensembles of numerical simulations. One such ensemble to be considered here has been developed by varying engineering parameters, such as those that change laser pulse shaping. This database will be called the macro- or engineering ensemble to distinguish it from micro-ensembles, discussed in Section 3.3, where the micro-physics models are adjusted.

The macro-ensemble parameter space will be defined by the parameters associated with shock timing the target. The shock timing is typically parameterized with 10 dimensions representing the speeds and launch times of four shocks and the strength and timing of the final accelerating pulse. The numerical simulations also capture spatial effects, such as the evolution of the implosion due to aspherical drive. The macro-ensemble also incorporates a three parameter variation of the shape of the driving radiation. This parameter space, small by NIF standards, is now 13-dimensional and is sampled at 2000 points. This sampling is as large as is practical given even the copious computing resources of LLNL. To sample the space requires running two asymmetric, or 2D simulations (with and without alpha deposition), and two spherically-symmetric, or 1D analogues of the 2D simulations. A single 2D simulation runs on 256 processors and requires more than 300 CPU-hours to complete. Thus, each 2000 sample ensemble requires more than 5 CPU-centuries to complete. This computation can be accomplished in a wall clock time of about one to two months. However, typically not all simulations run to completion since, at this scale, both hardware and software failures are common and computing resources are limited. In this case, the engineering ensemble discussed below consists of 1303 complete simulations. The battery of simulations gives spatiotemporally resolved data for the entire extent of each simulated implosion. This data is postprocessed by simulating the diagnostic instrument response to the implosion. When complete, a macro-ensemble provides simulations varying the 13 engineering or input parameters and an associated vector of postprocessed, distilled scalar diagnostic outputs to be discussed further in Section 5.

3.3 Micro-Simulations

During an experiment, different regions of an ICF capsule will access vastly different regions of the relevant space of physical parameters (such as the temperature, mass and electron densities, radiation field parameters, and others). It is then necessary for even simple simulations to describe plasma conditions that vary over many orders of magnitude; this in turn requires that the simulations contain good descriptions of a range of physics, from bulk thermodynamic properties to inherently quantum mechanical processes such as heat transport. During the evolution of the ICF experiment these "micro-physics" models become strongly interdependent and the output space of hydrodynamic simulations, when related to micro-physics models, can be expected to be quite complex.

For the most part, micro-physics modeling must be based on theoretical, first-principles calculations. These calculations will, naturally, be of varying (and often unknown) accuracy. The response of the simulation outputs to inaccuracies can be very complex, and a good understanding of this response is essential to interpreting the experimental data we observe. Attempts at elucidating the situation are normally approached by considering multiple models, often modified in *ad hoc* ways,

and comparing the models with specific aspects of the experimental data that are known to be sensitive. For example, the material equation of state (EOS), which relates sets of thermodynamic variables to one another, is known to depend in a complex way on the macro- and microscopic state of capsule materials and can be experimentally probed by observing the shock waves that are launched by the drive laser. An understanding of the topology of the space of shock parameters with respect to micro-physics could lead to an understanding of the errors in the underlying first principles EOS calculations, which could be used to motivate theoretical work and focused experimental campaigns. The complexity of ICF systems makes this process very difficult and there is ongoing interest in bringing advanced methods to bear.

In this work, we include a set of simulations created for the investigation of one particular micro-physics model. A set of physically motivated multipliers has been defined that changes important aspects of the atomic processes that govern the absorption and emission of X-rays in ICF capsules. This process is important in transporting energy, and it has been known for a long time that a good description requires detailed atomic physics calculations. This level of detail is incompatible with numerical constraints and so a compromise must be reached. The modifications to atomic physics we consider are designed to reflect important aspects of the radiation transport, and the simulation outputs have been chosen to allow direct comparison with experimental data. For this comparison, we use two metrics: a simple χ^2 distance between simulation and data, and a modified χ^2 that uses a Bayesian prior-predictive approach to include linear sensitivity to a large number of "engineering" noise sources. This modification has been developed in order to blur the line between macro- and micro-variables since the experimental observables (usually the same in both cases) clearly depend on both types simultaneously. A description of the Bayesian approach, and the simulations themselves, can be found in (Gaffney et al, 2013b) and (Gaffney et al, 2013a).

3.4 Current Analysis Efforts

Most efforts to analyze NIF simulation ensembles are guided by physical laws and reasoning. A central goal is to use physics models based on hydrodynamic and thermodynamic principles to generate a metric that quantitatively measures the distance from the well-performing plateau and summit, or more precisely, the cliff that surrounds it (Spears et al, 2012). Typically, the performance metric is assembled by making power law fits of yield to various implosion parameters, either input or driving quantities, or observable output quantities. Such efforts have given rise to useful metrics such as the theoretical ignition threshold factor (ITF) and the experimentally observable ignition threshold factor (ITFX). As an example, ITFX is defined by $ITFX = Y_n DSR^2$, where Y_n is the experimental yield from a nonburning fuel and DSR is an experimental measure of the quality of the fuel assembly. It is normalized so that the likelihood of ignition in simulation is 50 percent when $ITFX = 1$; this defines the location of the cliff (Fig. 1(b)).

Metrics such as ITFX appear to be satisfactory, ordering parameters for NIF yield performance from the perspective of measuring distance to a single cliff, but they are incapable of resolving any structure in parameter space other than the single, assumed plateau. Any more complicated structure is completely washed out. This paper will show that the presumption of a single, isolated maximum is overly simplifying and, in fact, broader analysis of macro-ensembles suggests the presence of at least two separated plateaus of nearly equal yield.

4 ND^{2}AV

This section will describe the target audience, the corresponding design choices, and the constraints of the system as well as the currently existing modules and implementation. Whereas ND^{2}AV initially has been conceived and designed specifically for the use in the NIC, it implements a general framework applicable to a large number of similar problems and tasks. The design and implementation of ND^{2}AV aims to strike the right balance between the needs of the target audience and the needs of the developers while respecting some practical constraints such as portability to compute clusters or even supercomputers.

4.1 Target Audience

Physicists and engineers working on the NIC are the initial target audience. Typically, these are highly trained and exceptionally capable individuals who, in many cases, have been involved in ICF-related problems for years or even decades. For most of this time reliable data has been extremely sparse and only recently simulations that approach realistic scenarios have become feasible. Therefore, progress has relied almost exclusively on the intuition and in-depth understanding of the physics involved to postulate unknown effects or interactions and ways to test these hypotheses. Advanced

data analysis and visualization have played an insignificant role in this process and as a result are considered experimental, unvetted, and thus suspect. Furthermore, given the decades of history in the corresponding fields, it is challenging to validate an unfamiliar approach beyond a reasonable doubt. Therefore, any analysis technique must, for the time being, work within the established frame of reference and in accordance with the physical intuition to have any practical impact. Furthermore, until such techniques have become more established, there exist few incentives for the scientists to invest time or resources in the development of such approaches, and thus, we consider even the experimental usage of ND^{2}AV by members of the NIC a significant success.

Currently, instead of a data-driven approach, scientists will postulate the existence of often highly complex, functional relationships that are then illustrated and validated with rather simple techniques such as scatter plots. A prime example is the definitions of performance metrics such as ITF and ITFX as discussed in Section 3.4 and Figure 1(b). In data analysis terms, the power law fit represents a complex, nonlinear embedding derived by hand, which transforms the data such that a simple scatter plot is sufficient to illustrate the desired dependencies. Conceptually, much of the analysis is driven by the search for functional relationships even in cases where such a relationship may not exist. For example, as discussed in Section 5.2, scientists will often (seemingly arbitrarily) split up the set of observations into a set of *drivers* and a set of outputs. The drivers are then considered *independent* variables and are used to explain/predict the set of *dependent* outcomes. This process reflects the belief that once the implosion has reached a certain state (i.e., high enough pressure, high enough velocity, etc.), the general outcome of a shot is predetermined.

used for serious analysis. Surprisingly, even comparatively standard approaches, such as nonlinear embeddings, e.g., Isomap, multidimensional scaling, etc., have similar problems. Even though an embedding into two dimensions looks like a scatter plot, the fact that the axes do not represent any one direction but rather a complex curve in high-dimensional space causes the results to be too abstract to be accepted easily. In light of these initial reactions, we have decided to focus less on new visualizations than on integrating more advanced analysis techniques into the one accepted visual aid, i.e., scatter plots. Nevertheless, we continue to believe that more advanced illustrations could provide significant advantages and new insights, as has been demonstrated for a host of other applications. To foster the adaptation process, we continue to integrate new visualization approaches to be used alongside more accepted techniques.

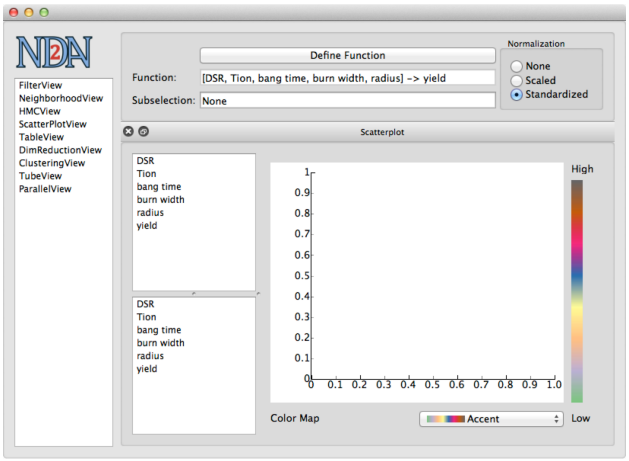


Fig. 2 The initial default window of ND^{2}AV containing a filter module to define, subselect, and scale the data and a scatter plot to display it. The panel on the left allows users to create new modules on demand by dragging them into the work area on the right.

4.2 Design Principles

As discussed in Section 2, there already exist a number of systems aimed at analyzing and visualizing the type of high-dimensional data of interest here. However, the needs and preferences of the target audience have led us to a very specific design, aimed at replicating the feel of the current approach while slowly introducing new techniques. The first consequence has been to focus primarily on scatter plots as a visual aid. We found that currently scientists have trouble connecting views such as parallel coordinates or topological structures to their intuitive understanding of physics. This situation makes it unlikely that such plots would be

Consequently, the initial view of ND^{2}AV is the filter module, shown in Figure 2, which provides convenient access to a scatter plot window as well as related functionalities such as loading files, normalizing and subselecting data, and defining derived quantities. The filter module allows a user to select data based on an arbitrary numpy (Oliphant, 2006) expression as well as to choose subsets of the given variables to use for analysis. Additionally, we allow users to enter an arbitrary expression to create a custom function, or to choose from some pre-loaded important indicator functions such as ITF (Haan et al, 2011) and ITFX (Spears et al, 2011) used extensively in the NIC. If desired, the module will automatically normalize the domain values to create

a common frame of reference by scaling each axis according to either its range or standard deviation. These features alone replicate the majority of existing functionalities in a more efficient and user-friendly manner than, for example, Excel, the current tool of choice.

For analysis, we focus on different approaches to cluster or more generally segment the data into coherent subsets with the goal of detecting localized dependencies and patterns in the resulting segmented scatter plots. ND^{2}AV provides a number of generic clustering techniques as well as topological segmentations, all of which are easily represented as color-mapped scatter plots. Surprisingly, the most abstract and least common option – topological segmentations – has received the most attention. The reason for such attention appears to be that general clustering, e.g., k-means, mean shift, etc., have the characteristic of a black box solution where the results are not easily related to the original input. Furthermore, choosing the various parameters, e.g., number of clusters, kernel bandwidth, etc., is somewhat arbitrary, especially to nonexpert users, yet drastically impacts the results. Instead, topological segmentations, especially Morse complexes (Edelsbrunner et al, 2003), have an intuitive and constructive definition: each segment represents all gradient paths ending at a maximum, i.e., a (high-dimensional) mountain. Additionally, the concept of a stable manifold or monotone cells is well aligned with the traditional viewpoint of describing data through functional relationships. Conceptually, a stable Morse cell can be thought of as, for example, a single Gaussian kernel centered around the maximum. Although likely not an accurate representation, this mental image appears to be more accessible and has led to a number of interesting observations.

Consequently, we have integrated the ability to define an arbitrary function into the initial filter view, allowing users to quickly define a *domain* and a *range* to construct a high-dimensional function that is subsequently analyzed. Nevertheless, topological decompositions also require setting a scale parameter, namely the *persistence* level (Edelsbrunner et al, 2002), to achieve optimal results. We present this choice as a curve displaying the number of features versus persistence, which in this context can be thought of as a complexity versus scale curve with which users are familiar. In particular, plateaus in this curve indicate a stable threshold, and typically even a non-expert can quickly determine which persistences correspond to noise and which to potentially interesting features.

The final design choice is how to represent and create complex workflows in an intuitive manner. One of the greatest advantages of ND^2 AV is the ability to quickly form and test new hypotheses and compare the results from different analysis techniques. For example, it is quite common to add or remove attributes from the domain or to explore the space using different attributes as the range. In such situations, users expect the entire analysis pipeline to be automatically reapplied to reflect any change, which on a system level requires the notion of workflows. However, the target audience is not familiar with such concepts and directly manipulating a workflow by, for example, explicitly connecting inputs and outputs of modules is impractical. At the same time, constructing a limited set of default workflows takes control away from the users and gives the appearance of a black box solution, making any result suspicious. To balance these requirements, we have divided the control over the workflows into two aspects.

The first, the user-control portion, is the ability to create additional modules and place them in a hierarchical fashion. As shown in Figure 2, the main window provides a list of available modules that the user can drag onto existing modules to instantiate the corresponding window. As discussed in more detail below, each filter module represents a *data context* and all modules within the same context are considered siblings and will be connected. By instantiating additional filter modules, the user can create a hierarchy of contexts to, for example, compare plots side by side. The second aspect of the workflow creation is forming the connections between modules within the same context. Even for simple workflows, these connections can quickly become rather complex (see Figure 4). To hide this complexity, the system currently forms all these connections fully and automatically. This design reduces the flexibility of the system to some extent. For example, there cannot exist two producers of the same data within the same context as connections will become ambiguous. However, we have not yet encountered a practical workflow where these restrictions have posed a problem, yet the automatic workflow creation greatly reduces the burden on the users.

Within each individual module, we rely as much as possible on default parameters, for example, as in clustering and dimensionality reduction modules. This choice certainly does not produce optimal results in all cases, but we have found that among the different algorithms available, there typically exists at least one for which the default results are quite acceptable. Furthermore, at the current stage most of these options are used only to corroborate results as the underlying techniques are not yet accepted as stable analysis tools. Therefore, in our experience, presenting users with different choices of techniques rather than an array of parameters to fine-tune any particular approach is more acceptable. Currently, the only "free" parameter not

linked to obvious user choices such as the number of clusters or segments to use is the number of neighbors. The reason is that the number of neighbors depends significantly on the inherent dimension of the data, the feature density, and the technique used, and thus picking useful defaults is nearly impossible. However, the system is typically fast enough to interactively adjust the number of neighbors, allowing users to test different choices and determine the sensitivity of the results with respect to change in the number of neighbors. By tightly coupling different analysis approaches in an interactive system, ND^{2}AV provides a simple-to-operate yet powerful analysis environment that has already led to multiple unexpected discoveries.

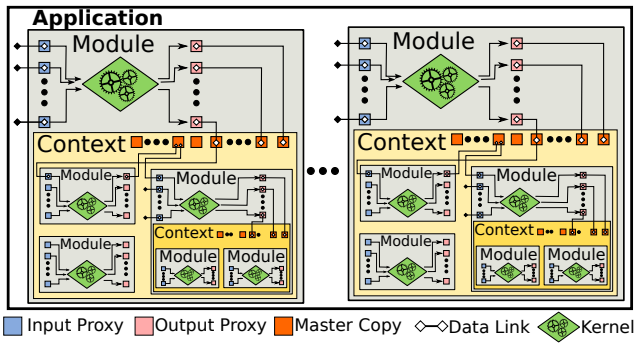


Fig. 3 Overview of the system design: The main application consists of (potentially multiple) modules each with their own data context. Submodules are placed hierarchically in their parent context. Each module declares input and output proxies that are dynamically linked to a single master copy of the corresponding data in the given context. Data can be linked across hierarchies through a trivial input-output connection.

4.3 System and Implementation

The two overarching goals from the development perspective are portability and flexibility. In this context, portability refers less to a wide variety of operating systems than to a broad set of computing environments. For example, in extreme cases the data for which ND^{2}AV has been developed may exist only on classified compute clusters or supercomputers, which cannot be expected to contain advanced graphics hardware or support the latest toolkits of one kind or another. As a result, we have chosen *Python* (van Rossum, 1995) as the lowest common denominator, enhanced where necessary by custom *C++* components for efficiency. Python provides easy access to a wide variety of libraries in machine learning, statistics, scientific computing, etc., and a low barrier of entry for new developers. The interface is based on PySide (PySide, 2010), a common Python implementation of the Qt system (Qt, 1995) that remains entirely decoupled from the base system and can be easily replaced if necessary.

ND^2 AV loosely follows a Model-View-Controller (Buschmann et al, 1996) paradigm in which all functionality is encapsulated in *modules* that are used by *views* to display and manipulate data. Instead of a traditional controller, our system uses the notion of a *Data Context* containing *Data Proxies* and *Shared Values*. A data context automatically links producers of data with consumers of data by maintaining a single master copy of the data to which each proxy and shared value contain a reference. Each time data changes, all registered proxies and shared values are notified.

Modules that contain a data context can contain submodules, which in turn can contain more contexts, creating a hierarchy of initially independent data contexts. Contexts are linked through input and output proxies of a given type. Whenever a module is created, it is assigned the proper context by either creating its own context or using the context of its parent. A module then dynamically declares the types of inputs it needs, the types of output it produces, and which values it would like to share with other modules, e.g., highlights colormaps, etc. The corresponding output proxies and shared values are immediately placed within the module's own context if it exists or into the parent's context otherwise, and they immediately become part of the corresponding data flow. Input proxies, however, are placed either into a parent context or if none exists are left isolated. In this manner, the input and output proxies of modules with internal contexts link the hierarchical contexts, see Fig. 3. The only difference between shared values and data proxies is that proxies carry version information, allowing modules to synchronize multiple inputs, and shared values are not linked across contexts by default. Currently, the only module that creates a context is the filter module. Its main purpose is to create and compare different functions or different filters applied to the same data. Within each filter module, all other modules are automatically and tightly coupled via the data context. Fig. 4 shows a typical configuration when comparing traditional clustering with topological techniques and the dataflow they imply. Note that the only user interaction required is the creation of modules through a drag and drop type interface and the placement of modules within the existing data contexts to determine the hierarchy of contexts. All other connections necessary to create the dataflow are handled automatically by attaching all necessary input, output, and shared data proxy within a data context to the appropriate master proxies of the context.

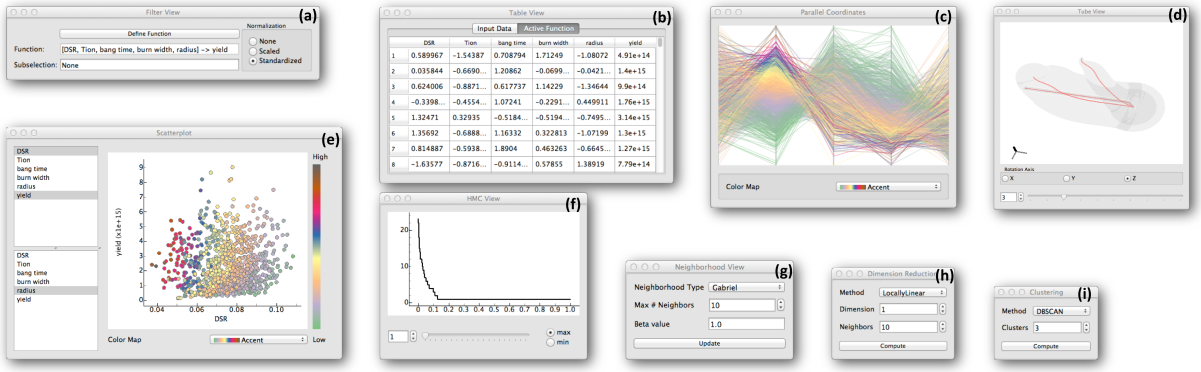


Fig. 5 Screen captures of all modules currently implemented in ND^{2}AV: (a) filtering and function creation; (b) table view; (c) parallel coordinate plots; (d) tube view showing center lines of the high-dimensional Morse complex embedded in 2D; (e) scatter plots; (f) hierarchical Morse complex and/or Morse-Smale complex view; (g) neighborhood graphs; (h) dimension reduction; (i) clustering.

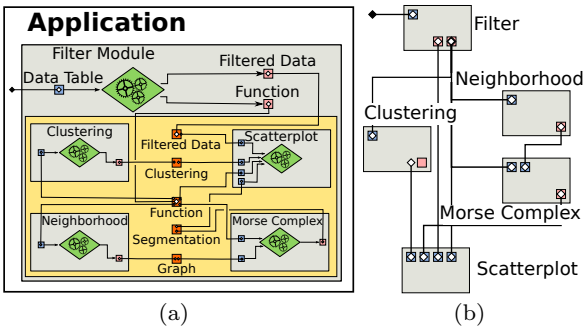


Fig. 4 An example dataflow comparing traditional and topological clustering. (a) The configuration of modules to realize this use case containing the central scatter plot alongside a clustering, neighborhood, and Morse complex module all linked via the data context. (b) The equivalent data flow drawn in a traditional manner.

One result of this design is that writing new modules for ND^{2}AV is simple yet it immediately places the module into a powerful, tightly coupled analysis framework. For example, Algorithm 1 shows the entire relevant code to make the *scikit-learn* (Pedregosa et al, 2011) clustering algorithms available in ND^{2}AV. The clustering module takes as input a function, which means n-dimensional points of a domain plus scalar function values, and outputs a clustering. The developer declares the corresponding input and output ports and links them through a specified function. Each time the input changes, the system will call the given function with the new input value and replace the output value with the return value of the callback function. In this case, the function simply calls the relevant method from *scikit-learn* and returns the result. The corresponding view creates the necessary buttons and interface elements to choose the clustering method and number of clusters desired and the module is ready to use. Even loosely spaced, with error checks, and comments, both files together total 104 lines of Python code.

```
class ClusteringModule(Module):
  def __init__(self,parent=None):
    self.makeInputPort("data",HDFunction)
    self.makeOutputPort("output",HDSegmentation)
    self.link(self.data,self.output,self.computeOutput)

  def computeOutput(self,data):
    if method == "DBSCAN"
      cluster = DBSCAN().fit_predict(data)
    elif method == "KMeans":
      cluster = KMeans(numClusters).fit_predict(data)
    elif method == "MeanShift":
      cluster = MeanShift().fit_predict(data)
    elif method == "Spectral":
      cluster =
      SpectralClustering(numClusters).fit_predict(data)
    return cluster
```

Algorithm 1: Code to make the *scikit-learn* clustering algorithms available in ND^{2}AV.

4.4 Existing Modules

Here we briefly describe the existing modules currently implemented in ND^{2}AV referring to Fig. 5 for the labeling. As briefly described in Section 4.2, the filtering and function creation module (a) allows users to load and filter data based on arbitrary expressions, to choose among existing indicator functions, to create custom function by specifying input and output variables, and to normalize the resulting functions according to either

their range or standard deviation. The scatter plots model (e), which is automatically synchronized with the filter module, reflects correlations between pairs of input variables with points colored by either the output values, or clustering and topological segmentation ids within each low-dimensional subspace. Both the table view (b) and parallel coordinate plots (Inselberg, 2009) (c) are dynamically linked with data points selected in the scatter plots, allowing pinpoint investigations of the data.

Eight types of neighborhood graphs (g) (e.g., approximated k-nearest neighbor graph, Delaunay triangulation, Gabriel graph, β-skeleton, as detailed in (Correa and Lindstrom, 2011)) can be chosen to connect points in a d-dimensional space. These graphs impose a combinatorial structure on the points and serve as basis for the topological analysis on the data, e.g., as showcased in the Hierarchical Morse(-Smale) complex view (f) and tube view (d) that encodes a topological summary. In (d) and (f), we adapt the techniques described in (Gerber et al, 2010) to compute the Morse complex of the chosen function. This algorithm provides a segmentation of the domain according to maxima (or minima) of the function or more intuitively a segmentation into high-dimensional mountains (or valleys). The corresponding view in (f) includes a modified persistence diagram called the persistence graph, first introduced in (Gerber et al, 2010), that allows users to choose topological segmentations of the function at different scales. Each extremum has an associate persistence that indicates at which *scale* this feature would be simplified and thus represents the significance of a feature. In particular, low persistence features are generally due to noise. This graph places the normalized persistence of the extrema (currently a user selects either minima or maxima) on the x-axis and the number of extrema that exist at that value on the y-axis. Furthermore, the shape of the graph indicates whether there exist stable ranges of scale to separate noise from features. The tube view (d) is similar to the visualization interfaces developed in (Gerber et al, 2010) and then enhanced in (Maljovec et al, 2013). The interface summarizes each Morse cell into a 1D curve using inverse linear regression in high-dimensional space, which is then projected onto a viewable 3D space. This view provides an atlas useful in understanding the connectivity between extrema, in the Morse-Smale case, areas of low or high sampling density, and also the relative size and shape of different topological segments, for example narrow peaks versus wide plateaus. Such an interface is easily extendable to combine and visualize any topological segmentations with appropriate regression techniques.

Finally, the clustering (i) and dimension reduction modules (h) explore existing techniques within *scikit-learn*, and previously introduced classic methods such as multidimensional scaling and spectral clustering.

5 Results

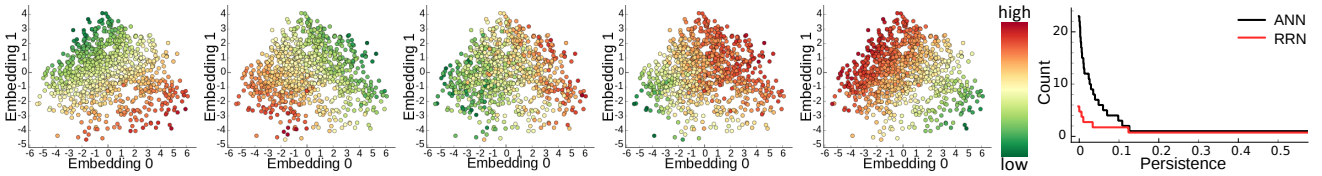
This section will present some of the initial results obtained using ND^{2}AV, some of which may challenge the basic assumptions of current models and intuitions. These results are especially valuable to our collaborators as they open new avenues for research and may ultimately lead to the formulation of new models and experiments.

5.1 Interactive Exploration

The most significant result achieved through using ND^{2}AV is a thorough and notably more comprehensive analysis of some of the data accumulated in the NIC. The ability to interactively and smoothly switch between analysis techniques, different figures of merit, different scales, and different datasets has enabled a rigorous exploration not feasible with existing tools. Therefore, even though most results could have been created through a combination of published (though not necessarily publicly available) techniques, only their combination into a common framework made them practically relevant.

Furthermore, while conceptually simple, the focus on segmented scatter plots has been instrumental in engaging scientists. In this context, making the seemingly abstract, topological segmentation easily available has proven very successful. The concept of a segmentation into regions around maxima or minima is intuitive, easy to explain and illustrate, and despite some of the unexpected results, well aligned with the physical intuition of scientists. In particular, comparing the clustering of the domain points with the topological decomposition of a function defined on this domain provides a new metric to judge how much structural information is due to the shape of the domains and how much is due to the functional relationship. This kind of insight provides important information on which parameters or variables are most significant in influencing the output.

On a systems level, only the highly modular design of ND^2 AV has allowed the visualization and analysis experts on our team to quickly include new tools suggested by previous results. For example, the discovery that some topological decompositions are closely related to the embedding (see Section 5.2) initially prompted the need to compare topological



(a) DSR (b) Tion (c) Bang Time (d) Burn Width (e) Radius (f)

Fig. 6 (a)-(e) 2D Isomap (Tenenbaum et al, 2000) embedding of the five driver quantities of the 1D engineering ensemble colored by axis. With few exceptions, the colormap suggests that the data is likely an embedded 2D manifold. (f) Persistence plot of the maxima of yield using a 10-nearest neighbor graph (black), indicating a large number of features yet no clear noise threshold. Instead using the relaxed relative neighborhood graph using 30 candidate edges (Correa and Lindstrom, 2011), (red) shows two strong and well-separated maxima.

decompositions against related clusterings, which as discussed above has been straightforward. Furthermore, as ND^{2}AV is becoming more widely used, we expect to continue to broaden its scope, for example, with different visualization techniques such as the Topological Spines (Correa et al, 2011) and additional analysis approaches.

Throughout the exploration we found a large number of unexpected and intriguing results, which have spawned many discussions and are currently being thoroughly investigated. Nevertheless, for brevity, we can present only some of the more interesting insights, all of which are the product of extensive exploration of the data. As such, the true value of ND^{2}AV is the ability to quickly perform a large number of different analysis steps, to experiment with parameters, and to correlate results. To highlight the process as well as the results, each of the subsections below walks through a specific example and some of the steps necessary to arrive at the final plot as well as some of the validation. Nevertheless, by necessity, only a small portion of the interactive process can be described in detail.

5.2 Engineering Simulations

The first two examples are ensembles of engineering or macro-physics simulations using a 1D or 2D model of the implosion, respectively (see Section 3.2). Note that, in this context, dimensionality refers to the physical space in which the simulations are performed, i.e., a 1D model represents a single ray connecting the center of the implosion with the outside. Similarly, a 2D model assumes a radially symmetric implosion computed in a 2D plane. These notations (e.g., 1D and 2D models) should not be confused with the number of input parameters or observables that are analyzed, which even for the simplest model can range from tens to hundreds. Typically, our collaborators create ensembles of simulation runs, each using slightly different input parameters typically varied according to a Latin-Hypercube (Tang, 1993) (LHC) design. From each run a large number of output quantities, such as peak velocity, yield, etc., are computed, and subsequently used to describe the resulting implosion. Furthermore, as discussed above, scientists will often organize the observed outcomes and inputs as *independent* and *dependent* variables, with the expectation that a (near) functional relationship explains the dependent variables as a function of the independent variables (also referred to as the *drivers*). In fact, most traditional analysis in the NIC is done by analyzing the dependencies of all observed outputs with respect to the drivers rather than analyzing the actual engineering inputs of the simulations. One important consequence is that since the drivers are in fact outputs of the simulations, they typically form a low-dimensional manifold embedded in some high-dimensional space rather than a space-filling sample created by an LHC design. This information must be taken into account when creating neighborhood graphs as, for example, the number of expected neighbors changes drastically according to the intrinsic dimension of the sampled space.

1D Engineering Ensemble. The first ensemble contains 1000 simulations from a 1D model, where each of the five drivers (down-scatter-fraction, ion-temperature, bang time, burn width, and capsule radius) was recorded alongside the overall yield of the implosion. Effectively, this dataset represents a high-dimensional distribution function of yield with respect to the chosen drivers. Therefore, the goal is to understand yield as a function of the drivers that form the domain. Following common practice in the NIC, the drivers are standardized according to their standard deviation to create a joint scale.

Since the drivers are products of the simulation (rather than part of the input LHC design), the first step is to guess the dimensionality of the domain. Fig. 6 shows a 2D Isomap embedding of the domain colored according to the five drivers. Except for a small number of outliers, each coordinate is linear in the embedding,

which strongly suggests that the domain is in fact a 2D submanifold embedded in five-dimensional space. In other words, the position of each point in the sample is fully described by any two drivers, and thus the underlying space is 2D (though not necessarily a flat plane). Note that these plots are not used to actually interpret the results, which as mentioned before is very difficult due to the non-linear transformations of the axis. Instead, they are only designed to provide a visual intuition that assuming an intrinsic dimension of two is reasonable. This in turn justifies a rather sparse (for a 5D space) k-nearest neighbor graph using 10 neighbors to perform a topological decomposition according to the maxima of yield following (Gerber et al, 2010). The resulting persistence plot (number of maxima versus persistence) is shown in Fig. 6(f) in black and indicates that there exist multiple significant maxima of yield, though no clear cut-off is visible. The lack of a "plateau" in this graph is troubling since it means there is no straightforward way to distinguish maxima due to noise or artifacts from true features. However, using one of the more complete and more stable graphs discussed in (Correa et al, 2011), such as a relaxed relative neighbor graph, removes some of the noise and indicates that there exist two primary maxima of yield as indicated by the red persistence plot of Fig. 6(f).

This result is significant in itself as it runs counter to the commonly held intuition that there exists a single plateau of high yield in the parameter space. Instead, it appears that (at least in this simulation ensemble) there exist two distinct regions of physics that can reach local maxima of yield. Furthermore, using the corresponding segmentation to color, traditional scatter plots reveal that the two resulting subsets show different dependencies among some of the variables. For example, Fig. 7(a) shows that whenever the red segment appears to exhibit virtually no correlation between the down-scatter-fraction and the yield, the blue segment suggests a strong positive correlation. Additional exploration shows that this result is consistent for different neighborhood graphs and number of neighbors.

Surprisingly, a very similar decomposition appears in a clustering of the domain points that does not consider the yield. Fig. 7(b) shows the result of a k-means clustering on the driver variables for two clusters. Furthermore, a mean shift clustering (Fig. 7(c)) splits the data into four clusters, each suggesting a different relationship between the down-scatter-fraction and yield. This figure indicates that most of the information about yield is already contained in the *shape* of the domain rather than in the actual values of yield. In other words, the decomposition into the two regions of physics leading to high yields can already be deduced from the domain, indicating that it is connected to the drivers rather than the resulting yield. This information has important implications for the pervasive practice of analyzing ICF through the drive versus output relationship, which will be discussed in more detail below.

2D Engineering Ensemble. This ensemble consists of two sets of 1303 simulations using a computationally much more expensive 2D model. For each simulation, 13 engineering parameters are varied according to an LHC design and 14 outputs with six specified as drivers are recorded. The two sets represent a deuterium-tritium (DT) target and a dudded tritium-hydrogen-deuterium (THD) target incapable of actually igniting. THD has the same density as a DT and is used to study the hydrodynamics in a diagnostic rich environment. Furthermore, 896 of the runs use a nominal "drive" (referring to the power and shape of the laser pulse used) whereas the rest use a reduced drive that can be seen as an additional binary input variable. This dataset is extremely rich and describing even a representative sampling of the potentially interesting aspects is beyond the scope of this paper. Instead, the discussion below focuses on one particularly interesting aspect that once again highlights discrepancies between the current thinking and the given data and how interactively linking different analysis tools led to these conclusions. The supplemental video shows the same analysis pipeline created dynamically and interactively.

Similar to the 1D engineering ensemble, the natural starting point of an analysis is to study how various outputs are related to the driver variables. In this case, six slightly different drivers are used: down-scatter-ratio, peak velocity, entropy, rhorba, pressure in the center, and hot spot radius. Just as before, a dimension reduction suggests that the six drivers form a 2D submanifold embedded in six dimensions leading us to choose a comparatively sparse graph with about 10 neighbors per sample. One interesting output is burn width, a measure of the width of the distribution of peak fusion reactivity. Its persistence graph of maxima (Fig. 8(a)) shows a strong plateau of three features with unusually high persistences indicating a very stable separation of scales.

The resulting segmentation according to maxima splits, for example, the scatter plot of burn width versus entropy (Fig.8(b)) contains three parts, each with its own slopes and behaviors. Particularly interesting is the fact that one would expect a positive correlation between the two variables as the larger entropy should lengthen the burn width. Instead, the largest (red) segment shows no such correlation, indicating that there exists another effect that increases burn width indepen-

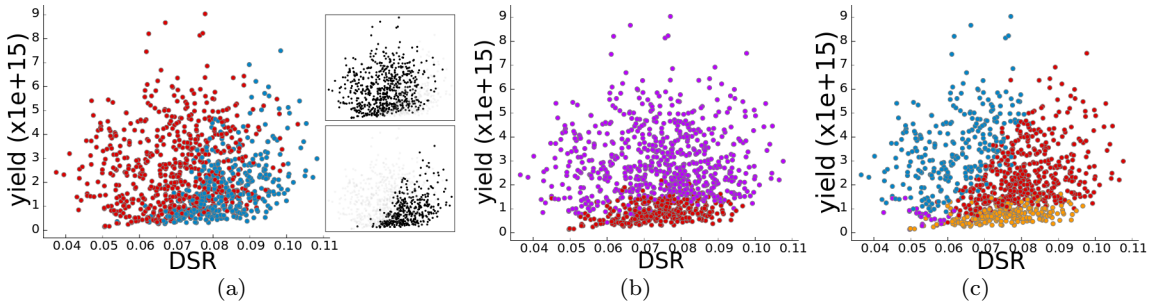


Fig. 7 (a) Topological decomposition corresponding to two maxima of the down-scatter-fraction (DSR)-to-yield scatter plot. The first segment (top right) shows little or no relationship between the two quantities, but the second segment (bottom right) suggests a strong positive correlation; (b) A k-means clustering with two clusters producing a decomposition similar to (a); (c) Mean shift clustering with default parameters producing four clusters, each suggesting a different correlation between DSR and yield.

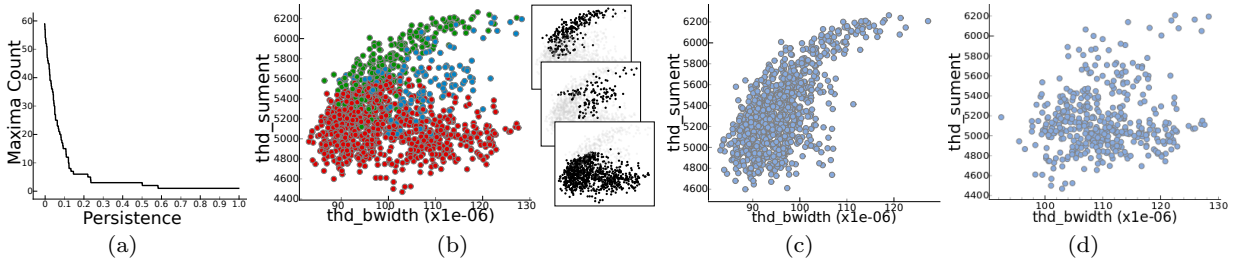


Fig. 8 (a) Persistence plot of the maxima of burn width as a function of the six drivers of the THD portion of the 2D engineering ensemble. (b) Entropy versus burn width plot of the same data colored according to maxima showing three different segments each with a different slope. (c), (d) Plots of the simulation using the nominal and reduced drive condition, respectively. Both subsets contain points in all three segments identified in (b).

dent of the entropy. This observation hints at another energy drain on the system that is of obvious interest.

One initial hypothesis for the cause for this split has been the differences between the nominal and the reduced drive. However, generating subsets of the scatter plots accordingly (Fig.8(a), (b)) disproves this notion. Both drives take part in two "branches," one with the expected burn width to entropy relation and one without.

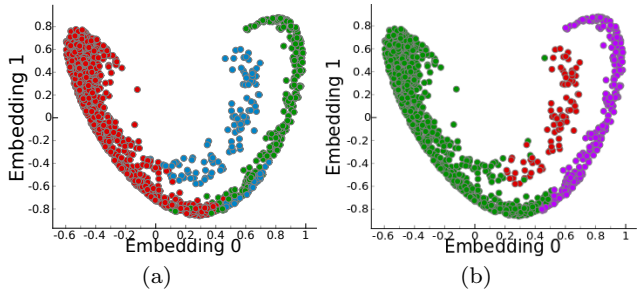


Fig. 9 Spectral embedding into two dimensions according to the drivers of the 2D engineering ensemble colored by the three dominant maxima of burn width (a) and according to a spectral clustering with three clusters (b). Both segmentations are similar, suggesting that burn width encodes little additional information not determined by the drivers.

Given the experience with the 1D ensemble, the next natural step is to consider the effects of the segmentation on the properly embedded data. Fig. 9(a) shows a spectral embedding (Ng et al, 2001) of the domain (i.e., using only the six drivers but not the burn width as input) color-coded according to maxima of burn width. The resulting segments are remarkably well separated in the embedding, again suggesting that the same segmentation could be achieved considering only the drivers. To explore this notion, Fig. 9(b) shows a spectral clustering into three clusters also based purely on the domain, which arguably produces an even better separation of the data. As before the non-linear embedding is only used to provide a more intuitive frame of reference to compare the clustering with the topological segmentation, not to interpret the results. The fact that both plots appear very similar suggests that burn width, or at least its topological structure, is more or less exactly predicted by the shape of the domain, with the actual function values adding little extra information. In other words, the the main information of interest is encoded in the relation of the simulation inputs to the drivers. However, analyzing either the drivers or burn width as a function of the 13 engineering parameters shows few structures, and none strong enough to

warrant consideration. This is likely due to the lack of sampling. As discussed above, the domain of the drivers form a 2D manifold for which 1303 points provide a reasonable sampling density. The same number of samples evenly distributed in the 13D input space, however, represents only about 1.7 samples per dimension, which appears to be too few for the complexity of the modeled behavior.

However, not all output quantities are well predicted by the shape. Unlike the 1D ensemble, for example, the overall yield as a function of the drivers also shows three significant maxima (Fig.10(e)) that are not obviously related to the embedding (Fig. 10(a)). Nevertheless, as shown in Fig. 10, the corresponding segmentation according to maxima produces interesting features in several outputs. All these features suggest that, against the common expectations, there exist multiple different regions in parameter space that locally maximize the yield. Each of these regions shows different correlations that may hint at slightly different underlying processes.

5.3 Micro-Physics Simulations

The final example consists of 7338 simulations runs of a 1D model concentrating on the micro-physics related to the X-ray absorption properties of the target (see Section 3.3). Five parameters related to X-ray absorption, emission, and the drive seen by the capsule have been varied according to a highly adaptive genetic algorithm aimed at producing simulations agreeing as close as possible with the experimental data. The output consists of 28 quantities of interest as well as two measures of distance to experimental data.

Apart from the adaptive nature of the sampling, all input parameters are equally varied, and thus, unsurprisingly, a much denser neighborhood is needed to construct even a connected graph. One of the first functions of interest is the distance to experiments and here we choose the modified χ^2 distance concentrating on the segmentation according to minima. The initial persistence plot shows a number of features with very low persistences (Fig. 11(b)). However, persistences are reported in percent of the overall range, which can be sensitive to outliers. In fact, the distribution of the two distance metrics shown in Fig. 11(a) shows a few outliers with extremely large distances that arise from the random initialization of the genetic algorithm and are not of interest. Consequently, we filter out simulations with $\chi^2 > 8000$, which has little effect on the shape of the persistence plot but, as expected, causes the persistences to increase (Fig. 11(b)). This observation suggests that these features are in fact significant relative to the bulk of the simulations.

Browsing through different combinations of inputs and outputs using scatter plots segmented according to the distance minima leads to the plot of the intensity of a nonthermal component added to the drive spectrum (drive-step-frequency) versus entropy shown in Fig. 11(c). It appears that different minima embody different linear slopes, suggesting subtly different physics might be active in the different portions of the parameter space. One possible explanation is that changes in any of the varied parameters result in poor timing of the shocks launched into the capsule, with resulting complex variations in the observed quantities.

6 Discussion

This case study demonstrates how important analyzing high-dimensional data can be to some of the most advanced areas of science. However, somewhat surprisingly, this impact did not come through the use of the most advanced visualization techniques or novel analysis approaches as one might expect. Instead, ND^{2}AV demonstrates that integrating the current workflow and enhancing well-established concepts rather than introducing new ones may result in better results and faster acceptance. As evident in Fig. 5, a few other modules are available and have been presented in ND^{2}AV, yet the unglamorous scatter plot has proved vital to any success. Arriving at this conclusion involved numerous discussions with the research scientists and engineers involved with the NIC. In these meetings, scientists repeatedly stressed their need to validate/invalidate their own hypotheses about the data without the confusion and uncertainty introduced by, for example, non-linear embeddings. More generally, we believe that at least in this application, a visual analysis approach of tightly integrating different techniques through most appropriate visual concepts is the best strategy.

Overall, ND^{2}AV has introduced significant new capabilities to NIC, one of the largest experimental efforts in the world. Over the course of a few months, this tool already has resulted in a number of unexpected results and new research directions. In particular, providing intuitive illustrations on how some commonly held assumptions may be incomplete or incorrect may ultimately lead to a better understanding of the differences between experiments and simulations, and high energy physics as a whole. Furthermore, ND^{2}AV presents an important stepping stone towards introducing more advanced analysis and visualization concepts to a new community with the potential for a significant theoretical and practical impact. In the context of NIC, the next steps are to integrate more advanced analysis capabili-

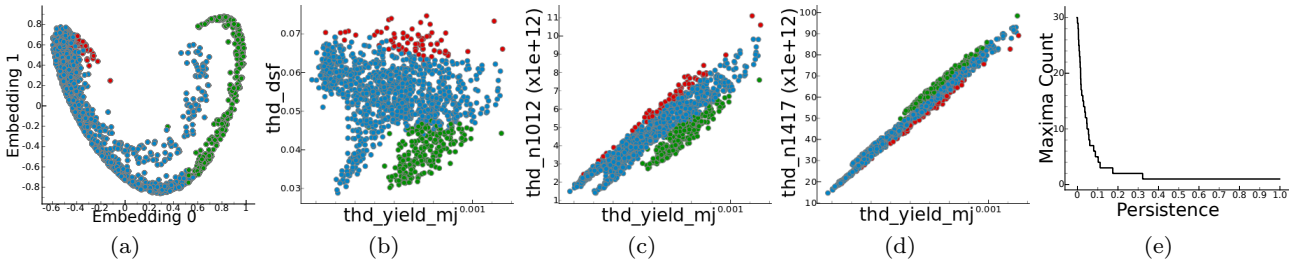


Fig. 10 (a) Spectral embedding into two dimensions according to the drivers of the 2D engineering ensemble colored by the three dominant maxima of yield. Unlike the burn width shown in Fig. 9(a), the topological structure of yield is not well predicted by the embedding. (b)-(d) Down-scatter-fraction, 1e10-1e12 neutron count, and 1e14-1e17 neutron count versus yield colored according to the maxima of yield. The persistence plot of yield maxima showing three well-separated maxima.

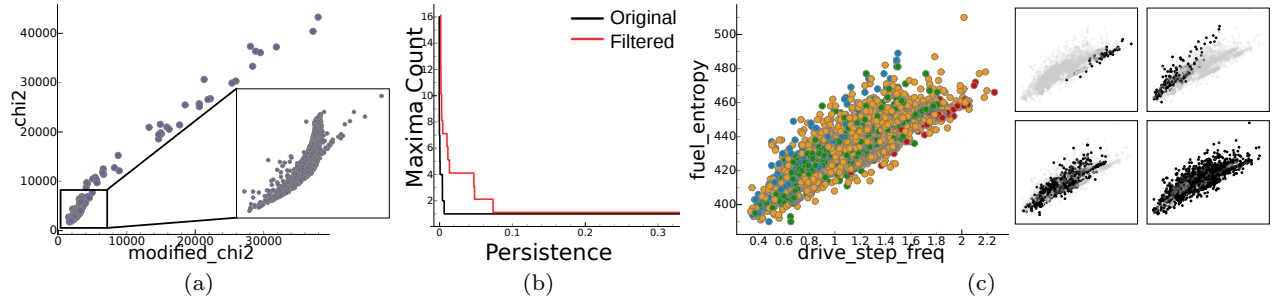


Fig. 11 (a) Scatter plot of χ^2 versus modified χ^2 distances of the micro-ensemble, indicating the presence of a small number of outliers with very large distances. The inset shows all points with $\chi^2 < 8000$. (b) Persistence plot of the minima of the modified χ^2 distance of the complete data (black) and only the samples with $\chi^2 < 8000$ (red). The basic structure is virtually the same but the (relative) persistences have increased as the global range has decreased. (c) Entropy versus drive-step-frequency colored according to the four well-separated minima of modified χ^2.

ties with the interactive pipeline such as, for example, comparisons with experimental data.

At the same time, our framework is generally applicable to a large number of research areas and its modular design makes tailoring ND²AV towards specific use cases easy. We are actively working with several other application areas such as climate analysis and nuclear reactor safety analysis to adapt and expand the tool. Going forward, we plan to release a version of ND²AV to the public domain in the hope of providing an easy-to-use platform to integrate and validate new techniques and tools.

Acknowledgements This work is performed in part under the auspices of the US DOE by LLNL under Contract DE-AC52-07NA27344, LLNL-CONF-658933, LLNL-JRNL-630732. This work is also supported in part by the NSF, DOE, NNSA, SDAV SciDAC Institute and PISTON, award numbers NSF 0904631, DE-EE0004449, DE-NA0002375, DE-SC0007446, DE-SC0010498, NSG IIS-1045032, NSF EFT ACI-0906379, DOE/NEUP 120341, DOE/Codesign P01180734.

References

Berger W, Piringer H, Filzmoser P, Gröller E (2011) Uncertainty-aware exploration of continuous parameter spaces using multivariate prediction. Computer Graphics Forum 30(3):911–920

Bergner S, Sedlmair M, Nabi S, Saad A, Möller T (2013) Paraglide: Interactive parameter space partitioning for computer simulations. IEEE Transactions on Visualization and Computer Graphics 19(9):1499–1512

Bertini E, Tatu A, Keim D (2011) Quality metrics in high-dimensional data visualization: An overview and systematization. IEEE Transactions on Visualization and Computer Graphics 17(12):2203–2212

Booshehrian M, Möller T, Peterman RM, Munzner T (2012) Vismon: Facilitating analysis of trade-offs, uncertainty, and sensitivity in fisheries management decision making. Computer Graphics Forum 31:1235–1244

Buschmann F, Meunier R, Rohnert H, Sommerlad P, Stal M (1996) Pattern-Oriented Software Architecture. Wiley

Chazal F, Guibas LJ, Oudot SY, Skraba P (2011) Persistence-based clustering in riemannian manifolds. Proceedings 27th Annual ACM Symposium on

Computational Geometry pp 97–106
Cook D, Swayne DF (2007) Interactive and Dynamic Graphics for Data Analysis: With Examples Using R and GGobi. Springer
Correa C, Bremer PT, Lindstrom P (2011) Topological spines: A structure-preserving visual representation of scalar fields. IEEE Transactions on Visualization and Computer Graphics 17(12):1842–1851
Correa CD, Lindstrom P (2011) Towards robust topology of sparsely sampled data. IEEE Transactions on Visualization and Computer Graphics 17(12):1852–1861
Edelsbrunner H, Letscher D, Zomorodian AJ (2002) Topological persistence and simplification. Discrete and Computational Geometry 28:511–533
Edelsbrunner H, Harer J, Zomorodian AJ (2003) Hierarchical Morse-Smale complexes for piecewise linear 2-manifolds. Discrete and Computational Geometry 30:87–107
Gaffney JA, Clark D, Sonnad V, Libby SB (2013a) Bayesian inference of inaccuracies in radiation transport physics from inertial confinement fusion experiments. High Energy Density Physics 9(3):457–461
Gaffney JA, Clark D, Sonnad V, Libby SB (2013b) Development of a bayesian method for the analysis of inertial confinement fusion experiments on the nif. Nuclear Fusion 53:073,032
Gerber S, Bremer PT, Pascucci V, Whitaker R (2010) Visual exploration of high dimensional scalar functions. IEEE Transactions on Visualization and Computer Graphics 16(6):1271–1280
Guo D (2003) Coordinating computational and visual approaches for interactive feature selection and multivariate clustering. Information Visualization 2(4):232–246
Haan SW, Lindl JD, Callahan DA, Clark DS, Salmonson JD, Hammel BA, Atherton LJ, Cook RC, Edwards MJ, Glenzer S, Hamza AV, Hatchett SP, Herrmann MC, Hinkel DE, Ho DD, Huang H, Jones OS, Kline J, Kyrala G, Landen OL, MacGowan BJ, Marinak MM, Meyerhofer DD, Milovich JL, Moreno KA, Moses EI, Munro DH, Nikroo A, Olson RE, Peterson K, Pollaine SM, Ralph JE, Robey HF, Spears BK, Springer PT, Suter LJ, Thomas CA, Town RP, Vesey R, Weber SV, Wilkens HL, Wilson DC (2011) Point design targets, specifications, and requirements for the 2010 ignition campaign on the national ignition facility. Physics of Plasmas 18(5)
Ingram S, Munzner T, Irvine V, Tory M, Bergner S, Möller T (2010) Dimstiller: Workflows for dimensional analysis and reduction. IEEE Conference on Visual Analytics Software and Technologies pp 3–10
Inselberg A (2009) Parallel Coordinates: Visual Multidimensional Geometry and Its Applications. Springer
Johansson S, Johansson J (2009) Interactive dimensionality reduction through user-defined combinations of quality metrics. IEEE Transactions on Visualization and Computer Graphics 15(6):993–1000
Kidder R (1974) Laser compression of matter: optical power and energy requirements. Nuclear Fusion 14(6)
Li JX (2004) Visualization of high dimensional data with relational perspective map. Information Visualization 3(1):49–59
Lindl J (1998) Inertial Confinement Fusion: The quest for ignition and energy gain using indirect drive. American Institute Of Physics
Lindl J, Atherton L, Amednt P, Batha S, Bell P, Berger R, Betti R, Bleuel D, Boehly T, Bradley D, Braun D, Callahan D, Celliers P, Cerjan C, Clark D, Collins G, Cook R, Dewald E, Divol L, Dixit S, Dzenitis E, Edwards M, Fair J, Fortner R, Frenje J, Glebov V, Glenzer S, Grim G, Haan S, Hamza A, Hammel B, Harding D, Hatchett S, Haynam C, Herrmann H, Herrmann M, Hicks D, Hinkel D, Ho D, Hoffman N, Huang H, Izumi N, Jacoby B, Jones O, Kalantar D, Kauffman R, Kilkenny J, Kirkwood R, Kline J, Knauer J, Koch J, Kozioziemski B, Kyrala G, Fortune KL, Landen O, Larson D, Lerche R, Pape SL, London R, MacGowan B, MacKinnon A, Malsbury T, Mapoles E, Marinak M, McKenty P, Meezan N, Meyerhofer D, Michel P, Milovich J, Moody J, Moran M, Moreno K, Moses E, Munro D, Nikroo A, Olson R, Parham T, Patterson R, Peterson K, Petrasso R, Pollaine S, Ralph J, Regan S, Robey H, Rosen M, Sacks R, Salmonson J, Sangster T, Sepke S, Schneider D, Schneider M, Shaw M, Spears B, Springer P, Stoeckl C, Suter L, Thomas C, Tommasini R, Town R, VanWonterghem B, Vesey R, Weber S, Wegner P, Widman K, Widmayer C, Wilke M, Wilkens H, Williams E, Wilson D, Young B (2011) Progress towards ignition on the national ignition facility. Nuclear Fusion 51(9)
Maljovec D, Wang B, Pascucci V, Bremer PT, Pernice M, Mandelli D, Nourgaliev R (2013) Exploration of high-dimensional scalar function for nuclear reactor safety analysis and visualization. Proceedings International Conference on Mathematics and Computational Methods Applied to Nuclear Science & Engineering pp 712–723
Matkovic K, Jelovic M, Juric J, Konyha Z, Gracanin D (2005) Interactive visual analysis end exploration of injection systems simulations. In: IEEE Visualization, pp 391–398
Munzner T (2009) A nested model for visualization design and validation. IEEE Transactions on Visualiza-

tion and Computer Graphics 15(6):921–928

Ng A, Jordan M, Weiss Y (2001) On spectral clustering: Analysis and an algorithm. In: Advances in Neural Information Processing, MIT Press, pp 849–856

Oliphant TE (2006) Guide to NumPy. Provo, UT, URL http://www.tramy.us/

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine Learning in Python . Journal of Machine Learning Research 12:2825–2830

Piringer H, Berger W, Krasser J (2010) Hypermoval: interactive visual validation of regression models for real-time simulation. Computer Graphics Forum 29(3):983–992

PySide (2010) Pyside. http://qt-project.org/wiki/PySideDocumentation

Qt (1995) Qt project. http://qt-project.org

R Development Core Team (2008) R: A language and environment for statistical computing. Vienna, Austria

Robey HF, Celliers PM, Kline JL, Mackinnon AJ, Boehly TR, Landen OL, Eggert JH, Hicks D, Le Pape S, Farley DR, Bowers MW, Krauter KG, Munro DH, Jones OS, Milovich JL, Clark D, Spears BK, Town RPJ, Haan SW, Dixit S, Schneider MB, Dewald EL, Widmann K, Moody JD, Döppner TD, Radousky HB, Nikroo A, Kroll JJ, Hamza AV, Horner JB, Bhandarkar SD, Dzenitis E, Alger E, Giraldez E, Castro C, Moreno K, Haynam C, LaFortune KN, Widmayer C, Shaw M, Jancaitis K, Parham T, Holunga DM, Walters CF, Haid B, Malsbury T, Trummer D, Coffee KR, Burr B, Berzins LV, Choate C, Brereton SJ, Azevedo S, Chandrasekaran H, Glenzer S, Caggiano JA, Knauer JP, Frenje JA, Casey DT, Gatu Johnson M, Séguin FH, Young BK, Edwards MJ, Van Wonterghem BM, Kilkenny J, MacGowan BJ, Atherton J, Lindl JD, Meyerhofer DD, Moses E (2012) Precision shock tuning on the national ignition facility. Physical Review Letters 108

van Rossum G (1995) Python tutorial. Tech. Rep. CS-R9526, Centrum voor Wiskunde en Informatica (CWI)

Seo J, Shneiderman B (2005) A rank-by-feature framework for interactive exploration of multidimensional data. Information Visualization 4(2):99–113

Singh G, Mémoli F, Carlsson G (2007) Topological methods for the analysis of high dimensional data sets and 3D object recognition. Eurographics Symposium on Point-Based Graphics pp 91–100

Spears B, Brandon S, Clark D, Cerjan C, Edwards J, Landen O, Lindl J, Haan S, Hatchett S, Salmonson J, Springer P, Weber S, Wilson D (2011) The experimental plan for cryogenic layered target implosions on the National Ignition Facility—the inertial confinement approach to fusion. Physics of Plasmas 18(5)

Spears BK, Glenzer S, Edwards MJ, Brandon S, Clark D, Town R, Cerjan C, Dylla-Spears R, Mapoles E, Munro D, Salmonson J, Sepke S, Weber S, Hatchett S, Haan S, Springer P, Moses E, Kline J, Kyrala G, Wilson D (2012) Performance metrics for inertial confinement fusion implosions: Aspects of the technical framework for measuring progress in the national ignition campaign. Physics of Plasmas 19(5)

Sutherland P, Rossini A, Lumley T, Lewin-Koh N, Dickerson J, Cox Z, Cook D (2000) Orca: A visualization toolkit for high-dimensional data. Journal of Computational and Graphical Statistics 9(3):509–529

Tang B (1993) Orthogonal array-based latin hypercubes. Journal of the American Statistical Association 88(424)

Tatu A, Albuquerque G, Eisemann M, Schneidewind J, Theisel H, Magnor M, Keim D (2009) Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. IEEE Symposium on Visual Analytics Science and Technology pp 59–66

Tenenbaum JB, de Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290(5500):2319–2323

Theus M, Urbanek S (2008) Interactive Graphics for Data Analysis: Principles and Examples (Computer Science and Data Analysis). Chapman & Hall/CRC

Torsney-Weir T, Saad A, Moller T, Hege HC, Weber B, Verbavatz JM (2011) Tuner: Principled parameter finding for image segmentation algorithms using visual response surface exploration. IEEE Transactions on Visualization and Computer Graphics 17(12):1892–1901

VisuMap Technologies Inc (2009) VisuMap - a high dimensional data visualizer (visumap white paper). Calgary, Alberta

Ward MO (1994) Xmdvtool: integrating multiple methods for visualizing multivariate data. Proceedings IEEE Conference on Visualization pp 326–333

Waser J, Fuchs R, Ribicic H, Schindler B, Bloschl G, Groller M (2010) World lines. IEEE Transactions on Visualization and Computer Graphics 16(6):1458–1467

van Wijk JJ, van Liere R (1993) Hyperslice: visualization of scalar functions of many variables. Proceedings IEEE conference on Visualization pp 119–125