Variable Interactions in Query-Driven Visualization

Luke J. Gosink, *Student Member, IEEE*, John C. Anderson, *Student Member, IEEE*,
E. Wes Bethel, *Member, IEEE*, and Kenneth I. Joy, *Member, IEEE*

Abstract—Our ability to generate ever-larger, increasingly-complex data, has established the need for scalable methods that identify, and provide insight into, important variable trends and interactions. Query-driven methods are among the small subset of techniques that are able to address both large *and* highly complex datasets. This paper presents a new method that increases the utility of query-driven techniques by visually conveying statistical information about the trends that exist *between* variables in a query. In this method, correlation fields, created between pairs of variables, are used with the cumulative distribution functions of variables expressed in a user's query. This integrated use of cumulative distribution functions and correlation fields visually reveals, with respect to the solution space of the query, statistically important interactions between any three variables, and allows for trends between these variables to be readily identified. We demonstrate our method by analyzing interactions between variables in two flame-front simulations.

Index Terms—Multivariate Data, Query-Driven Visualization

1 Introduction

Obstacles hindering scientific research may be broadly categorized into two separate but overlapping groups. The first category, concerned mainly with issues of throughput, includes the challenges inherent to efficiently managing and visualizing large-scale datasets. The second category includes the difficulties associated with attaining insight from datasets of high-complexity.

Query-driven visualization (QDV) is well-suited for performing analysis and visualization on datasets that are both large *and* highly complex [21, 22]. Tools like FastBit leverage highly efficient (in terms of speed and compression) data management techniques to rapidly identify and visualize "regions of interest" within a dataset [10, 15, 16, 17]. Specified as Boolean range queries (e.g., $(Methane < 0.25\text{M})$ AND $(760\text{ Torr} \leq Pressure \leq 820\text{ Torr})$), these regions of interest tend to be significantly smaller subsets of the original dataset; thus, these regions require less time and computational effort to analyze, visualize, and interpret.

Well-characterized range queries are capable of identifying spatial regions where many domain-specific events occur: combustion flame fronts, vortices, chemical reaction fronts, etc. Beyond indicating these regions, however, queries reveal little about variable interactions or complex trends that lie in the domain of these characterizations. In such regions of interest, it is the behavioral trends *between* variables, or groups of variables, that are more important in providing insight than the traits or locations of individual variables alone. The challenge is to extend the strengths of QDV with methods that identify behavioral trends and provide insight into regions of interest through coherent and meaningful visualizations.

The novel contributions of this work are techniques that extend the capabilities of QDV by providing intuitive insight in determining:

- how sets of variables in complex datasets interact throughout regions of interest, and
- the role other variables play in influencing these interactions.

We utilize the cumulative distribution functions (CDFs) of all variables in a query, to reveal initial information about statistical regions of interest within the query's solution space. The CDF for each variable is formed by integrating over the query's solution space and accumulating the variable's values as a histogram. Statistically, the solution set of a query is represented as an aggregate of histograms, one histogram for each variable expressed in the query.

We extend this analysis further by incorporating the use of correlation fields. Correlation fields provide insight into *localized* correlation between any two variables. By mapping a correlation field onto a third variable's isosurfaces (specifically, the statistically important isovalues suggested by the variable's CDF), statistically important interactions between any three variables in a dataset are readily visualized, allowing for trends between variables in a user's query to be identified.

In this method, CDFs and correlation fields are constrained to the query's *solution space*. By working exclusively in the query's solution space, this method takes full advantage of the performance benefits inherent to QDV strategies. Specifically, computational efforts are only focused on regions that have been rapidly identified (via a query engine) as "interesting" by the user's query. This method's integrated analysis extends current query solutions by revealing statistical trends of interactivity (i.e., dependency and independence) between any triad of variables in the solution space of the query.

We discuss research related to our efforts in the next section. Section 3 defines the terminology used to describe multivariate queries. We utilize this terminology, in Section 4, to describe our method. In Section 5, we present a detailed discussion of variable interactions in the context of flame-front analysis. Concluding remarks are given in Section 6.

2 Previous Work

Numerous methods have been developed for the management and visualization of large datasets. Some of the more effective software solutions include *local adaptive mesh refinement* (AMR) [4], and the bitvector compression schemes used in query based strategies [22].

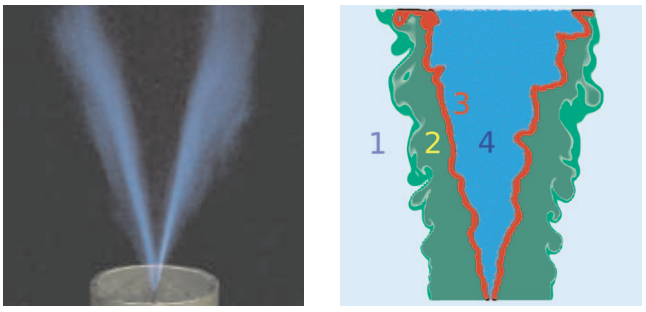
Based on adaptive numerical discretization techniques used to approximate solutions for partial differential equations (PDE), AMR algorithms work on the principle that many physical phenomenon (e.g., second-order systems such as wave equations, particles interfaces, radiative heat transfer, charged-fluid plasma models, low-Mach combustion, etc.) exhibit variations in scale [4]. Standard numerical solutions require the discretization of the phenomenon's domain space to be within an acceptable margin of error (as determined by the algebraic analogues of the PDE). In many cases, due to the need for high resolution (i.e., due to discontinuities), the discretization becomes very dense and costly.

Rather than creating a uniform grid of high density, AMR begins

Luke J. Gosink, John C. Anderson, and Kenneth I. Joy are with the Institute for Data Analysis and Visualization (IDAV) at the University of California, Davis. E-mail: ljgosink@ucdavis.edu, janderson@ucdavis.edu, and kijoy@ucdavis.edu.

E. Wes Bethel is with the Computing Sciences Division and Scientific Visualization Group at Lawrence Berkeley National Laboratory, E-mail: ewbethel@lbl.gov.

Manuscript received 31 March 2007; accepted 1 August 2007; posted online 27 October 2007. Published 14 September 2007.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.



(a) Photographed Flame (b) Illustrated Flame

Fig. 1. These images show photographed (a) and illustrated (b) methane combustion. The illustration, at right, also labels the principle regions of the methane flame involved in the combustion process. The outer region (1) corresponds to the atmospheric environment, around the methane fuel source, consisting of nitrogen, oxygen, and small amounts of water. The mixing region (2) contains the fuel source and thus the highest concentration of methane. The reaction region (3) is where the "flame fronts" and a diverse collection of chemical intermediates reside. The product region (4) contains the highest concentration of water and carbon dioxide – the final products in the combustion process.

with a relatively course grid and adaptively refines the mesh resolution *only* in regions requiring higher levels of accuracy. The range of resolution is usually fixed with an assigned initial coarse resolution. Regions requiring further detail are refined recursively until either the maximum level of refinement is reached, or the desired level of accuracy has been achieved. For a single simulation, numerous levels of refinement are possible. Advantages of AMR include dramatically reduced storage requirements and computational analysis (on a per-node basis).

Strategies based upon QDV are designed on the premise that smaller subsets of data are usually the genesis of insight or breakthroughs to new trends [2, 11]. The goal of query-driven analysis is the rapid isolation and visualization of "interesting data" (as determined by user dictated Boolean range queries) from the dataset. One of the most daunting challenges in data analysis is the rapid identification and retrieval of these records. Efficient compressed bitmap index technologies, like FastBit [21, 22], achieve this goal through compression and accelerated algorithms that allow for the rapid retrieval of queried records.

The field of multivariate visualization has also been actively researched. Scatterplot matrices are perhaps the most direct visualization of correlation between multiple variables [20]. The primary limitation of scatterplot matrices lies in the fact that scatterplots show correlations in data range space, but do not offer the ability to see such correlations in the native spatial coordinates of the dataset.

Methods for determining correlation have been well studied, and there are many approaches established for measuring correlation between pairs of scalar variables [12]. The use of scalar field gradients as a comparison measure of correlation [7] was recently used in work that utilized Multifield-graphs [14]. In that work, correlation between sets of variables (i.e., fields) was visualized with a graph structure.

Parallel coordinates are a compact, two-dimensional visualization of trends between multiple variables [13]. The concept of parallel coordinates was extended by Fua et al. [9] to support large datasets, clustering, and interactive brushing. Blended texture approaches for two-dimensional scalar fields have also been explored with the goal of compactly conveying as much information as possible about variable trends on a single surface [18, 19].

While two-dimensional visualizations of multivariate correlation are common, three-dimensional and time-varying visualizations are comparatively rare. Akiba et al. [1] have performed initial work in the area of higher dimensional multivariate rendering by utilizing a parallel coordinates-based transfer function widget to direct the rendering of multivariate and time-varying datasets.

There is still a need for methods, applicable to large multivariate data, that elucidate how variables behave and interact within regions of interest. In this paper, we combine CDFs and correlation fields to extend the solutions generated by QDV approaches. With this method we are able to evaluate (with respect to the query's solution space) any three variables, and the interactions between these variables, in a meaningful way.

3 Cumulative Distribution Functions and Correlation Fields

In this section we formalize two principle statistical concepts used in our method. The first portion of this section defines the CDF construction performed for each variable used in a given query. The second part of this section explains the process for creating correlation fields between two variables. Correlation fields are used, with the CDFs of the query's variables, to visually evaluate any three variables, and the interactions between these variables, within the solution space of the query.

3.1 The Univariate Universe

Let a finite universe $\mathscr{U}$ be defined such that $\mathscr{U} \subsetneq \mathbb{R}^3$. Additionally, let $\mathscr{U}$ be defined to be univariate; there exists exactly one mapping function, $\mathbf{f}$, considered valid on $\mathscr{U}$ that maps each of $\mathscr{U}$'s elements $\mathbf{f} : \mathbb{R}^3 \rightarrow \mathbb{R}$. This is the usual definition for a scalar field dataset in $\mathbb{R}^3$ space.

A common notion in such a universe is that of the level set. In $\mathscr{U}$, level sets are defined implicitly through the mapping function $\mathbf{f}$. Thus, the explicit level set for any real value $\mathbf{a}$ is equivalent to the exhaustive solution of the *inverted* mapping function when evaluated at $\mathbf{a}$ (i.e., $\mathbf{f}^{-1}(\mathbf{a})$). Notationally, $\mathscr{S}_{\mathbf{f}\rightarrow\mathbf{a}}$ denotes the explicit level set formed for the real value $\mathbf{a}$ by the mapping function $\mathbf{f}$ in the universe $\mathscr{U}$:

$$\mathscr{S}_{\mathbf{f}\rightarrow\mathbf{a}} = \{\mathbf{x} : (\mathbf{x} \in \mathscr{U}) \wedge (\mathbf{f} : \mathbf{x} \rightarrow \mathbf{a})\} \tag{1}$$

When $\mathscr{U} \subsetneq \mathbb{R}^2$ or $\mathscr{U} \subsetneq \mathbb{R}^3$, these level sets are used to construct isocurves and isosurfaces respectively.

Without loss of generality, we assume that a finite number of level sets $(\mathscr{S}_{\mathbf{f}\rightarrow\mathbf{a}}, \ldots, \mathscr{S}_{\mathbf{f}\rightarrow\mathbf{n}})$ are generated by $\mathbf{f}$, and that the cardinality of these sets is also finite. We observe that the mapping function $\mathbf{f}$ may then be represented by the discrete random variable $\mathscr{X}_{\mathbf{f}}$. This random variable takes any real value $(\mathbf{a}, \ldots, \mathbf{n})$ from the possible functional mappings of $\mathbf{f}$, and thus describes the distribution behavior of the real-valued function $\mathbf{f}$ over the sample space of $\mathscr{U}$. Appropriately, a *probability mass function* (*PMF*) for $\mathscr{X}_{\mathbf{f}}$, denoted $p_{\mathscr{X}_{\mathbf{f}}}$, may be constructed:

$$p_{\mathscr{X}_{\mathbf{f}}}(\mathbf{a}) = P(\mathscr{X}_{\mathbf{f}} = \mathbf{a}) = \frac{\|\mathscr{S}_{\mathbf{f}\rightarrow\mathbf{a}}\|}{\|\mathscr{U}\|} \tag{2}$$

Here $P(\mathscr{X}_{\mathbf{f}} = \mathbf{a})$, or more generally P(**E**), denotes the probability of event **E** occurring. Specifically, Equation 2 states the probability that the mapping function $\mathbf{f}$ will map a randomly selected element in $\mathscr{U}$ to the real value $\mathbf{a}$. Combining all *PMF*s for a single random variable forms the appropriate *CDF*, denoted $F_{\mathscr{X}_{\mathbf{f}}}$, for $\mathscr{X}_{\mathbf{f}}$:

$$F_{\mathscr{X}_{\mathbf{f}}}(\mathbf{n}) = P(\mathscr{X}_{\mathbf{f}} \leq \mathbf{n}) = \frac{\|\bigcup_{i=\mathbf{a}}^{i=\mathbf{n}} \mathscr{S}_{\mathbf{f}\rightarrow\mathbf{i}}\|}{\|\mathscr{U}\|} \tag{3}$$

Here Equation 3 states the probability that the mapping function $\mathbf{f}$ will map a randomly selected element in $\mathscr{U}$ to any real value between $\mathbf{a}$ and $\mathbf{n}$ (inclusive).

3.2 The Multivariate Universe

We extend this definition of a random variable to the case of multivariate universes where there exists multiple valid mapping functions $(\mathbf{f}, \mathbf{g}, \ldots, \mathbf{k})$ that map each of $\mathscr{U}$'s elements $(\mathbf{f}, \mathbf{g}, \ldots, \mathbf{k} : \mathbb{R}^3 \rightarrow \mathbb{R})$. As with the univariate universe, a unique random variable, $\mathscr{X}_{\mathbf{f}}, \mathscr{X}_{\mathbf{g}}, \ldots, \mathscr{X}_{\mathbf{k}}$, may be created for each valid mapping function that

describes the distribution behavior of this real-valued function over the sample space of $\mathscr{U}$.

For each of these random variables, a *joint PMF* and *CDF* may be defined. For example, given real constraints for random variable $\mathscr{X}_\mathbf{g}$ in a multivariate universe $\mathscr{U}$, the joint mass and distribution functions of $\mathscr{X}_\mathbf{f}$ are expressed:

$$p_{\mathscr{X}_\mathbf{f}|\mathscr{X}_\mathbf{g}}(\mathbf{a}|\mathbf{c} \leq \mathscr{X}_\mathbf{g} \leq \mathbf{d}) = \frac{\|\mathscr{S}_{\mathbf{f}\rightarrow\mathbf{a}} \cap \bigcup_{i=\mathbf{c}}^{i=\mathbf{d}} \mathscr{S}_{\mathbf{g}\rightarrow\mathbf{i}}\|}{\|\mathscr{U}\|} \quad (4)$$

$$F_{\mathscr{X}_\mathbf{f}|\mathscr{X}_\mathbf{g}}(\mathbf{n}|\mathbf{c} \leq \mathscr{X}_\mathbf{g} \leq \mathbf{d}) = \frac{\|\bigcup_{i=\mathbf{a}}^{i=\mathbf{n}} \mathscr{S}_{\mathbf{f}\rightarrow\mathbf{i}} \cap \bigcup_{i=\mathbf{c}}^{i=\mathbf{d}} \mathscr{S}_{\mathbf{g}\rightarrow\mathbf{i}}\|}{\|\mathscr{U}\|} \quad (5)$$

Observe that Equations 4 and 5 above may be extended such that the conditioning of random variable $\mathscr{X}_\mathbf{f}$ includes any additional random variables ($\mathscr{X}_\mathbf{h}, \ldots, \mathscr{X}_\mathbf{k}$) that define distributions of valid mapping functions on $\mathscr{U}$.

3.3 Multivariate Queries

We define a Boolean query, $\mathbf{q}$, as a composite function that operates on those mapping functions considered valid on $\mathscr{U}$. More formally, $\mathbf{q} : (\mathbf{f} : \mathbb{R}^3 \rightarrow \mathbb{R}) \rightarrow \mathbb{B}$, resulting in the solution set for the query, denoted as $\mathscr{Q}$:

$$\mathscr{Q} = \{\mathbf{x} : (\mathbf{x} \in \mathscr{U}) \wedge (\mathbf{q} : (\mathbf{f} : \mathbf{x} \rightarrow \mathbb{R}) \rightarrow 1)\} \quad (6)$$

Note that if $\mathscr{U}$ is multivariate, the query function may extend to include any of the possible valid mapping functions for $\mathscr{U}$.

The solution set $\mathscr{Q}$ is a reduced sample space of $\mathscr{U}$. Appropriately, the level sets valid in $\mathscr{Q}$ for each mapping function will also be reduced (to some degree) from their unconditioned counterparts. These reduced level sets visually represent the conditional reduction imposed by the constraints of the composite query function $\mathbf{q}$. For our purposes we restrict such general query functions to be Boolean range queries (e.g., "$X < 100$, $X > 10$, or $0 < X < 10$") where continuous regions of $\mathscr{U}$ are defined to be in $\mathscr{Q}$.

3.4 Correlation Fields

The differential operator $\bigtriangledown$ acts as a composite function that can take any mapping function valid on $\mathscr{U}$ to construct a gradient field in $\mathbb{R}^3$.

Given two mapping functions (for illustrative purposes, $\mathbf{f}$ and $\mathbf{g}$), represented by respective random variables $\mathscr{X}_\mathbf{f}$ and $\mathscr{X}_\mathbf{g}$, a scalar field indicating the localized correlation measured between these two variables may be constructed by observing the cosine of the angle between the two corresponding gradient fields generated by $\bigtriangledown\mathbf{f}$ and $\bigtriangledown\mathbf{g}$. In our method, this measurement of correlation is taken by observing the *normalized* dot product between the two gradient fields. Using random variables $\mathscr{X}_\mathbf{f}$ and $\mathscr{X}_\mathbf{g}$, the correlation field is represented:

$$\mathbf{C}_{\mathscr{X}_\mathbf{f}\mathscr{X}_\mathbf{g}} = \bigtriangledown\mathbf{f} \cdot \bigtriangledown\mathbf{g} \quad (7)$$

Correlation coefficients (being summary statistics) and scatterplots are only indicators of "global" trends. Correlation fields, on the other hand, are useful for the spatial and "local" analysis of correlation; they identify regions, restricted to a particular query's solution set, where important variable interactions and trends take place.

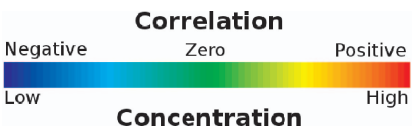


Fig. 2. Colormap used for pseudocoloring both correlation and scalar concentration fields. Blue corresponds to negative correlation within correlation fields, green to very little or no correlation, and red to positive correlation. For scalar concentrations, the lowest concentration values map to blue, while higher values map to red.

4 Method

In our method, we utilize CDFs in conjunction with correlation fields to elucidate how variables in a given query interact. Given the solution set $\mathscr{Q}$ for a specific Boolean range query in a multivariate universe, we conduct our analysis in the following manner. We first begin by determining the individual "contributions" to $\mathscr{Q}$ made by each variable in the query. This is performed by integrating over the reduced sample space of $\mathscr{Q}$, and recording each variable's values in a histogram (one histogram per variable).

Next, these variables are classified based upon their respective histograms. This classication is performed by sorting the entries in each variable's histogram by frequency of occurrence in $\mathscr{Q}$. From the definitions in Sections 3.1 and 3.2, we observe that this is equivalent to defining each variable's distribution in $\mathscr{Q}$ as a random variable, and sorting all possible values for these random variables based upon probability (as expressed in Equation 4). Additionally, we observe that these histogram entries are represented by level sets, and thus isosurfaces, in $\mathscr{Q}$. Continuing with this notion of level sets, the following general classification is performed:

- Level sets for a given variable with a probability greater than one standard deviation above the mean of the variable's other level set probabilities are labeled as *principle level sets*. Principle level sets make up the majority of the query's solution. These level sets offer the most potential for gaining insight, when used in conjunction with a correlation field, as they convey the most representative behavior of the variable in the reduced sample space of $\mathscr{Q}$.

- All other level sets are *secondary level sets*.

After determining the principle level sets, any two variables of interest are selected to construct a correlation field – performed by observing the *normalized* dot product between two generated gradient fields as shown in Equation 7. Note that the chosen variables do not have to be part of the query, merely constrained by it. Working exclusively in the query's solution space allows for this method to take full advantage of performance benefits inherent to QDV strategies. Specifically, computational efforts are only focused on regions that have been rapidly identified (via a query engine) as "interesting" by the user's query.

Principle level sets are then rendered through this correlation field as isosurfaces and "colored" with the transfer function corresponding to the correlation field generated from the two previously selected variables. This method of exploration offers several levels of insight. Consider two variables $\mathscr{X}_\mathbf{f}$ and $\mathscr{X}_\mathbf{g}$. Let $\mathbf{C}_{\mathscr{X}_\mathbf{f}\mathscr{X}_\mathbf{g}}$ be the correlation field between $\mathscr{X}_\mathbf{f}$ and $\mathscr{X}_\mathbf{g}$. To explore the interaction of a third variable $\mathscr{X}_\mathbf{h}$ with $\mathbf{C}_{\mathscr{X}_\mathbf{f}\mathscr{X}_\mathbf{g}}$, we render isosurfaces (from respective principal level sets) of $\mathscr{X}_\mathbf{h}$ colored by scalar values from $\mathbf{C}_{\mathscr{X}_\mathbf{f}\mathscr{X}_\mathbf{g}}$. Below, we discuss two behaviors that may be observed in this type of visualization (additional behaviors are addressed in Section 5):

- **Behavior:** Isosurfaces of $\mathscr{X}_\mathbf{h}$ are consistently multi-colored, with strong positive, zero, *and* strong negative correlation intermixed across the surface.
  OR
  Isosurfaces of $\mathscr{X}_\mathbf{h}$ are predominantly and consistently colored by *either* strong positive, zero, *or* strong negative correlation.

  Interpretation: This type of behavior indicates little correlation between $\mathscr{X}_\mathbf{h}$, and the correlation field $\mathbf{C}_{\mathscr{X}_\mathbf{f}\mathscr{X}_\mathbf{g}}$. If varying $\mathscr{X}_\mathbf{h}$ consistently shows *either* no change in the correlation between $\mathscr{X}_\mathbf{f}$ and $\mathscr{X}_\mathbf{g}$, *or* "random" changes, it is difficult to infer any relationship between $\mathscr{X}_\mathbf{h}$ and $\mathbf{C}_{\mathscr{X}_\mathbf{f}\mathscr{X}_\mathbf{g}}$.

- **Behavior:** Isosurfaces of $\mathscr{X}_\mathbf{h}$ are colored predominantly with a single correlation value (e.g., strong positive, zero, *or* strong negative). As the isosurface value changes, the predominant correlation value coloring the surface also changes.

  Interpretation: This behavior indicates the possible existence of a scientifically-important relationship between $\mathscr{X}_\mathbf{h}$ and $\mathbf{C}_{\mathscr{X}_\mathbf{f}\mathscr{X}_\mathbf{g}}$.

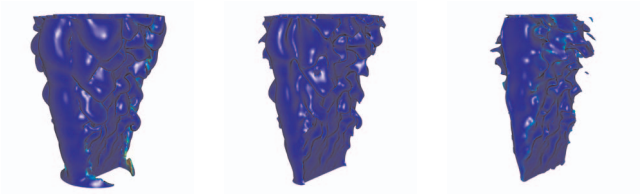


Fig. 3. From the methane combustion dataset, temperature isosurface values of 300, 1000, and 1800 degrees Celsius are shown "colored" with the correlation field derived from oxygen (O_2) and carbon dioxide (CO_2) (see Figure 4(c)). Note that the isotherms lie in regions of strong negative correlation, regardless of temperature value; this is visually evident because each surface is predominantly colored blue (negative correlation from Figure 2). Physically, this indicates that any correlation between O_2 and CO_2 (shown in Figure 4(c)) is independent of temperature.

One explanation of this behavior may be that variable $\mathscr{X}_{\mathbf{h}}$ influences the correlation between $\mathscr{X}_{\mathbf{f}}$ and $\mathscr{X}_{\mathbf{g}}$. Another explanation may be that the correlation between $\mathscr{X}_{\mathbf{f}}$ and $\mathscr{X}_{\mathbf{g}}$ effects the value of the variable $\mathscr{X}_{\mathbf{h}}$.

Using this visualization method, it is instructive to look for isosurface regions corresponding to near-zero correlation. Unlike areas of positive or negative correlation (low entropy, high mutual information), near-zero correlation regions correspond to areas of high entropy and low mutual information. The Intermediate Value Theorem may be used to show that to move from an area of positive correlation to an area of negative correlation, or *vice versa*, one must pass through a point of zero correlation. As with any system, the corresponding increase in entropy must be caused by external events or internal processes. If large parts of an isosurface of one variable coherently follow such a transformation between two other variables (positive to negative, or negative to positive), then it is likely that the isosurface rendered at the point of near-zero correlation between the two other variables is important to the external event or internal process causing the entropy change. This method's utility is its ability to identify the transitions (i.e., areas of maximal entropy) that indicate regions of important change or interactivity between variables within the user's query.

5 Application and Analysis

We apply the techniques developed in Sections 3 and 4 to two combustion datasets. The first dataset models the combustion of a methane flame, while the second dataset models the turbulent combustion of an ultra-lean premixed hydrogen flame.

In the study of combustion, it is important to understand flame anatomy. Figure 1 shows photographed and illustrated methane combustion. In the illustrated image, important regions of the flame are labelled:

1. The *outer region*, in light blue, corresponds to the atmospheric environment, found around the fuel source, consisting of nitrogen, oxygen, and trace amounts of water.

2. The *mixture region*, in green, corresponds to the region in which the fuel source, and thus the highest concentration of methane, exists. Note that in this region, oxygen and other atmospheric gasses also still exist.

3. The *reaction region*, in red, is the most complex region chemically. In this region, "flame fronts" arise along the leading fronts of fuel and oxygen combustion.

4. The *product region*, in blue, contains the highest concentration of both water and carbon dioxide – the final products of methane combustion.

The above information also generalizes to hydrogen combustion flames. In both methane and hydrogen combustion mechanisms, intermediate compounds and radical species are produced and consumed in the *reaction region*. The numerous interactions between these species, and the more stable chemical species (i.e., the hydrogen or methane fuel, oxygen, nitrogen, etc.), form a sequence of events we define as combustion. In the remainder of this section we explore the methane and hydrogen combustion datasets, using our method to identify scientifically interesting relationships between variables.

5.1 Methane Combustion

We discuss in detail the application and resulting analysis of our method to a methane combustion dataset[1]. The abbreviated, general chemical equation for methane combustion is shown in Equation 8 below:

Figures 5 and 6

$$2O_2 + CH_4 \xrightarrow{C_2H_4} 2H_2O + CO_2 + \lambda \quad (8)$$

Figures 3 and 4

Here, oxygen (O_2) and methane (CH_4) react to produce water (H_2O), carbon dioxide (CO_2), and energy (λ). Ethylene (C_2H_4), one of the intermediate chemical species produced and consumed in the combustion process, is shown above the red loop.

Our analysis focuses on isosurfaces of temperature (i.e., isotherms) in respect to two separate correlation fields: the correlation field derived from O_2 and CO_2, and the correlation field derived from H_2O and C_2H_4.

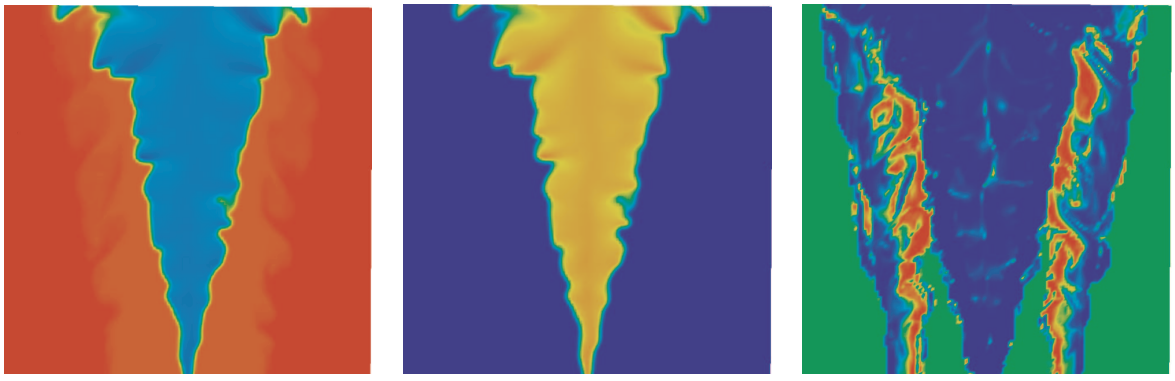
To begin, Figure 4 depicts the concentrations of O_2 and CO_2 ((a) and (b), respectively). In these images, red regions indicate areas of high concentrations and blue regions indicate areas of low concentrations. Figure 4(c) shows a 2-dimensional slice of the correlation field derived from O_2 and CO_2. For this image, red regions indicate strong positive correlation, blue regions indicate strong negative correlation, and green regions indicate regions where there is little to no correlation.

Note that areas of strong positive correlation (visible in the mixing regions) in Figure 4(c), which are not discernable in either Figure 4(a) or (b), are a result of using *normalized* dot products to construct the correlation field. Normalization allows for the observation of interactions between these gasses, even at trace concentration levels.

To interpret the correlation image in Figure 4(c), observe that in the blue regions (i.e., when moving to the center of the image from either just left or right of center), the concentrations of CO_2 and O_2 are simultaneously increasing and decreasing respectively; in red regions, O_2 concentrations increase in the same direction as CO_2. In the green regions, because they are uncorrelated, no discernible trend exists between O_2 and CO_2. Observing Figure 1(b), it is clear that the region where O_2 and CO_2 are positively correlated in Figure 4(c) corresponds to the mixture region. Figure 4(c) shows strong negative correlation outside the mixture region, especially on the mixture region's outer edge, and in the whole of the product region. This negative correlation in the product region is intuitive from the inverse relationship shared between these two chemical species as defined in Equation 8 (i.e., O_2 is a reactant and CO_2 is a product).

Coloring the isosurfaces of a single variable with the correlation field derived from two other variables is useful in identifying the important interactions occurring, or trends shared, between all three variables (as discussed in Section 4). Figure 3 depicts isosurfaces of increasing temperature that are colored with the correlation field shown

[1]The simulated dataset used in this paper was generated from the DRM-19 subset of the GRI-Mech 1.2 methane combustion mechanism [8] for chemical kinetics [3, 6] which considers over 20 chemical species and 84 fundamental reactions. Over 325 reactions and 53 species may be obtained by using the GRI-Mech 3.0 mechanism.

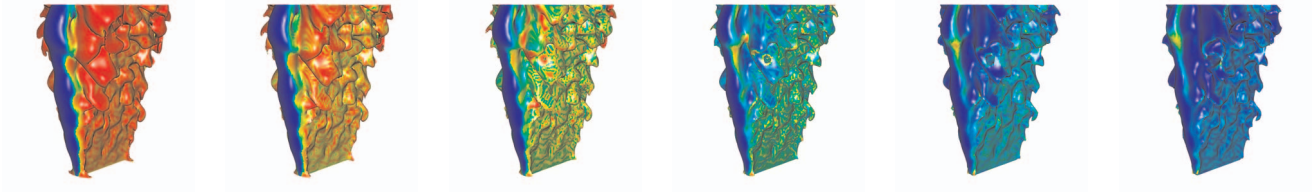


(a) Oxygen (O_2) concentrations (b) Carbon Dioxide (CO_2) concentrations (c) Correlation space for O_2 and CO_2

Fig. 4. Oxygen (O_2) and carbon dioxide (CO_2) concentrations from the methane combustion dataset are shown in (a) and (b), respectively. The derived correlation field for these two gasses is shown in (c). Areas of strong positive correlation (shown in red) in the mixing regions in (c), which do not appear in either (a) or (b), are a result of using *normalized* dot products to construct the correlation field. Normalization allows for the observation of interactions between these gasses, even at trace concentration levels.

(a) Water (H_2O) concentrations (b) Ethylene (C_2H_4) concentrations (c) Correlation space for H_2O and C_2H_4

Fig. 5. Water (H_2O) and ethylene (C_2H_4) concentrations from the methane combustion dataset are shown in (a) and (b), respectively. The derived correlation field for these two gasses is shown in (c). The switch from strong positive correlation to strong negative correlation in the reaction region corresponds to the area in which C_2H_4 is both produced and consumed, and H_2O is produced, in the process of combustion. The strong correlation (both positive and negative) in the center of the flame, as well as the atmospheric region, demonstrates the correlation field's ability to show fine-scale interactions (see also Figure 4(c)).



(a) (b) (c) (d) (e) (f)

Fig. 6. These images depict increasing ((a) through (f)) isosurface values of temperature (isotherms) colored by values of the correlation field derived from water (H_2O) and ethylene (C_2H_4) (see Figure 5(c)). As temperature values increase, the predominant correlation between H_2O and C_2H_4 along the isotherms shifts from strongly positive (red) in (a), to strongly negative (blue) in (f). This shift suggests that temperature is itself negatively correlated with the H_2O-C_2H_4 correlation.

in Figure 4(c). All isosurfaces in Figure 3 (shown at 300, 1000, and 1800 degrees Celsius) lie in regions of strong negative correlation, regardless of temperature value; this is visually evident because each surface is predominantly colored blue (negative correlation from Figure 2). This indicates that temperature has little influence on the relationship between O_2 and CO_2, or conversely, that any interactions between O_2 and CO_2 are independent of temperature (i.e., they are neither exothermic or endothermic which indicates that these two species are NOT reactive in this simulation).

As a second example in this dataset, we examine the relationship between the stable species H_2O, which exists in the product region, and the intermediate species C_2H_4, which exists primarily in the reaction region. As previously mentioned, interactions between intermediate species and more stable species are essential for combustion to occur. Figure 5 shows concentrations of H_2O and C_2H_4 ((a) and (b) respectively), and the corresponding correlation field generated between these two variables in (c). Interestingly, the switch from strong positive correlation to strong negative correlation in the reaction region corresponds to the area in which ethylene is both produced and consumed, and water is produced, in the process of combustion. Areas of strong correlation (both positive and negative) in the center of the flame, as well as in the atmospheric region, demonstrate the correlation field's ability to show fine-scale trends (see also Figure 4(c)).

As with our first example, it is possible to color isotherms with values in the H_2O-C_2H_4 correlation field as shown in Figure 6. Unlike Figure 3, which showed little correlation between temperature and the O_2-CO_2 correlation, temperature seems to be correlated (to some degree) with the relationship between H_2O and C_2H_4. Visually, this is supported by Figure 6. At low temperature, Figure 6(a), the isotherm is in the mixture region. Along the isotherm in the concentration fields, the gradients point inward, thus yielding strong positive correlation in the correlation field. This correlation means that the direction of greatest increasing H_2O concentration is very similar to the direction of greatest increasing C_2H_4 concentration.

As the temperature value increases, the isosurface moves steadily into the reaction region. Along these isotherms the correlation between H_2O and C_2H_4 begins to change. By the third isotherm (Figure 6(c)) there is little correlation between the chemical species (as per Figure 2, green indicates regions of near-zero correlation). Thus, over the isotherm, one can infer very little about changes in H_2O concentration from changes in C_2H_4 concentration (and *vice versa*).

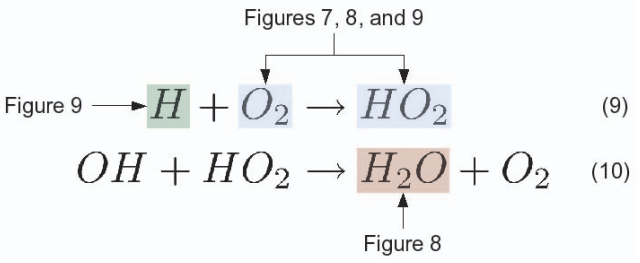
It is in these regions of near-zero correlation where we make the following observation. As discussed in Section 4, those regions where the correlation field for two variables is uniformly minimal over a predominant region on the isosurface of a third variable indicate areas of important changes or interactivity between all three variables. Additionally, if large parts of an isosurface of one variable coherently follow a transformation in correlation between two other variables (positive to negative, or negative to positive) as depicted in Figure 6(c), then it is likely that the isosurface rendered at the point of near-zero correlation between the two other variables is important to the external event or internal process causing the entropy change.

To interpret this in the context of our dataset, we know that C_2H_4 is a temporary species formed and consumed in the process of combustion. In the flame front regions, we expect the concentration of such species to be relatively high. Conversely, we expect H_2O concentrations to be relatively low on the edge of the flame front. The correlation between C_2H_4 and H_2O will then be minimal, i.e., higher entropically, in these areas where the production and consumption of C_2H_4, and production of H_2O is greatest (i.e., the regions at, or near, the flame fronts).

Increasing the temperature further reveals regions of strong negative correlation, specifically in the product region. We know from basic knowledge about the regions (see Figures 5(a) and (b)), that concentrations of H_2O are low in the mixture and outer region (where the temperature is also low), thus the correlation field indicates that C_2H_4 concentrations should also be low. In the product region, where the temperature is hottest, we know that concentrations of H_2O are high so that from the correlation field we then know that concentrations of C_2H_4 should be very low. This pattern suggests that temperature itself is *negatively* correlated with the correlation between H_2O and C_2H_4.

5.2 Hydrogen Combustion

In this section, we discuss the application and resulting analysis of our method to a hydrogen combustion dataset[2]. The general chemical equation for hydrogen combustion is: $O_2 + 2H_2 \Rightarrow 2H_2O + \lambda$. Rather than focusing on the abridged, general reaction, we instead analyze two intermediate reactions principle to the formation of water in combustion:



Equation 9 shows a hydrogen radical (H) interacting with elemental oxygen (O_2) to produce a perhydroxyl radical (HO_2). Equation 10 shows a hydroxyl radical (OH) reacting with HO_2 to produce water (H_2O) and O_2.

In the hydrogen combustion dataset, isosurfaces of increasing H_2O and H concentrations are separately rendered through a correlation field constructed from O_2 and HO_2.

Figures 7(a) and 7(b) show the concentrations of O_2 and HO_2, respectively. Here, as with the methane example, red regions indicate the highest concentrations of a given species and blue indicates the regions of lowest concentrations. Figure 7(c) depicts a 2-dimensional slice of the correlation field generated between the gradients of these two chemical species. In this image, red regions indicate strong positive correlation, blue regions indicate strong negative correlation and green regions indicate regions where there is little to no correlation.

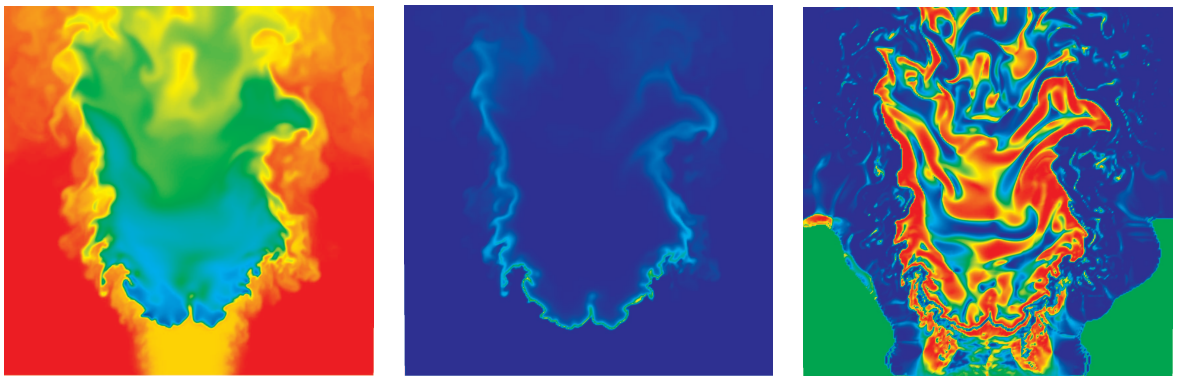
Figure 8 depicts isosurfaces of increasing concentrations of H_2O that have been colored by values from the correlation field derived from O_2 and HO_2 (see Figure 7(c)). These rendered isosurfaces exhibit "striations" (i.e., bands of negative, zero, and positive correlation) in the correlation field. With increasing concentrations of H, correlation is shown to increase within each striation. This behavior suggests that H concentration is positively correlated with the O_2-HO_2 correlation.

A possible explanation for the positive correlations, exhibited across striations of negative, zero, and positive correlation, is that in the hydrogen dataset, unlike the methane dataset, burning occurs unevenly along the isotherms. Such variations in combustion influences both the rates of reactions and the locations of reaction fronts. As such, transitions in correlation are expected to occur at different concentrations in the isosurfaces of H_2O (as Figure 8 seems to depict).

Observe the areas of highest entropy (green) on isosurfaces of medium H_2O concentration in Figures 8(b), (c), and (d). These areas indicate the regions in which the reaction from Equation 10 is most likely to occur. Note that for higher-concentration H_2O isosurfaces (Figures 8(e) and (f)), entropy has decreased, corresponding to high positive correlation between O_2 and HO_2.

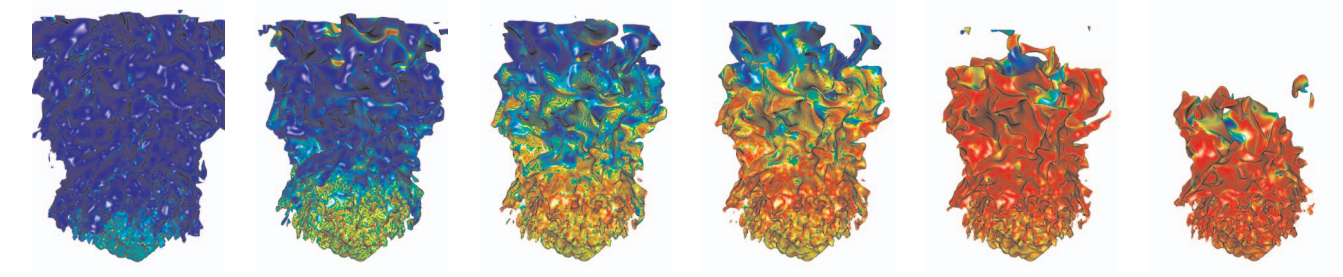
These observations suggest that areas of high H_2O production are not necessarily areas of high H_2O concentration; this indicates that H_2O is being moved by some means away from the primary reaction area. Alternatively, this may suggest that there are additional H_2O-producing reactions (beside that shown in Equation 10) driving the production of water. Having made this observation, it would be instructive to visually explore additional correlation fields, constructed from species involved in *other* H_2O-producing reactions.

[2]As with methane combustion, this chemical reaction is a simplification. Our simulated hydrogen combustion dataset is based on the NEW and DER mechanisms [5] which consist of 12 chemical species involved in over 34 reactions.



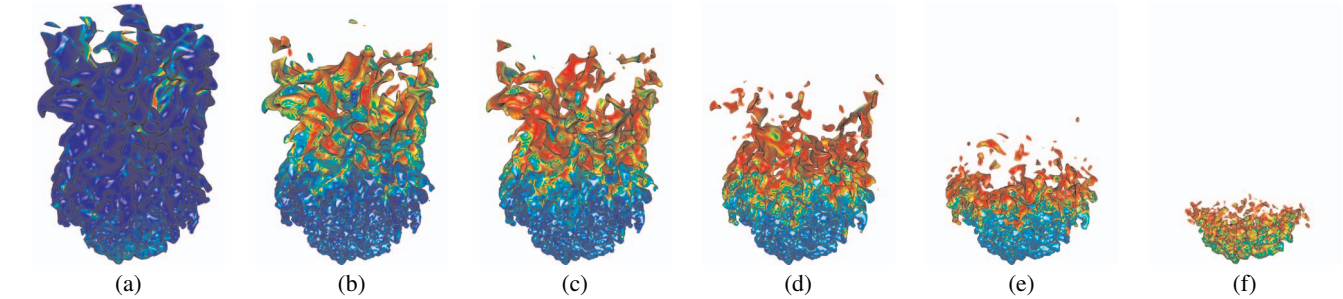
(a) Oxygen (O_2) concentrations (b) Perhydroxyl (HO_2) concentrations (c) Correlation field for O_2 and HO_2

Fig. 7. We show slices through an ultra-lean premixed hydrogen flame combustion dataset. Unlike the methane combustion dataset, in which burning occurs "evenly" along isotherms, combustion in the hydrogen dataset is very uneven and results in a characteristic "bubbling" shape. Oxygen (O_2) and perhydroxl radical (HO_2) concentrations are shown in (a) and (b), respectively. The derived correlation field for these two gasses, which highlights their turbulent interplay within the reaction region, is shown in (c).



(a) (b) (c) (d) (e) (f)

Fig. 8. These images depict increasing ((a) through (f)) iso-concentrations of water (H_2O) colored by values from the correlation field of oxygen (O_2) and perhydroxyl radical (HO_2) (see Figure 7(c)). As water concentration increases, the predominant correlation along the isosurface shifts from strongly negative (blue) in (a), to strongly positive (red) in (f). This shift suggests that H_2O concentration is itself positively correlated with the O_2-HO_2 correlation. Local variations in this observed correlation (e.g., the bottom of the isosurfaces transitioning from negative correlation to positive correlation faster than the top of the isosurfaces) are likely due to the fact that in the hydrogen dataset, unlike the methane dataset, burning occurs unevenly along the isotherms. Such variations in combustion influences both the rates of reactions and the locations of reaction fronts. As such, transitions in correlation are expected to occur at different concentrations in the isosurfaces of H_2O (as this image depicts). In Section 5.2, we hypothesize that these higher-concentration H_2O isosurfaces (which correspond to high positive-correlation regions between O_2 and HO_2) indicate that there are additional H_2O-producing reactions (beside that shown in Equation 10) driving the production of water.



(a) (b) (c) (d) (e) (f)

Fig. 9. These images depict increasing ((a) through (f)) iso-concentrations of hydrogen radicals (H) colored by values from the correlation field of oxygen (O_2) and perhydroxyl radical (HO_2) (see Figure 7(c)). Each isosurface exhibits striations in the correlation field (i.e., bands of negative, zero, and positive correlation), and as H concentration increases, correlation increases within each striation (i.e., negative correlation tends to become positive). This behavior suggests that H concentration is itself positively correlated with the O_2-HO_2 correlation. Image (f) indicates the simultaneous existence of near-zero correlation (i.e., high entropy) *and* high H concentrations; in Section 5.2 we hypothesize that this combination is a driving force in the reaction shown in Equation 9.

In Figure 9, isosurfaces of increasing H concentration are shown. Isosurfaces of H, similar to H_2O, highlight striations in the correlation field that increase in correlation within each striation as the concentration of H increases. These changes in correlation (from negative to positive), that coincide with the increasing concentrations of H, suggest a positive correlation between H concentrations and the O_2-HO_2 correlation.

Additionally, there is one significant difference between H and H_2O in reference to O_2-HO_2 correlation. Figure 8 suggests that the reaction in Equation 10 is unable to drive H_2O concentrations to their highest values. In Figure 9(f), however, we see that the highest concentrations of H correspond to regions of high entropy (i.e., near-zero correlation). This is significant because the simultaneous existence of high entropy with high concentrations of H suggests that (f) highlights a region in which Equation 9 consumes significant amounts of H and produces a significant amount of HO_2.

6 Conclusion

We have presented a method that increases the utility of QDV techniques. By mapping a correlation field onto the isosurfaces determined by CDFs, we can visualize interactions between any three variables, allowing trends between variables in a user's query to be identified. The set of experiments we have presented demonstrates the flexibility of our approach in its ability to:

- hold a correlation field constant and explore the effects that varying isosurfaces may have (as we have shown with the hydrogen data set), and
- hold an isosurface variable constant and observe its interactions in varying correlation fields (as we have shown with the methane dataset).

Query systems have the potential to solve many large-scale visualization problems. In applications where the scientist only needs to examine a portion of the dataset, or where the features of the data are contained in a small region, query systems give a method by which limited portions of a dataset can be accessed. Visualization techniques that operate on these queries have tremendous potential to provide insight and yield scientific breakthroughs with datasets, that up until now, we could not have previously addressed.

Future work will focus on extending our method in two principle areas. One important extension we are actively exploring is the ability to visualize correlation fields that represent relationships between multiple variables. Identifying correlation between level sets colored by the values from such fields, and the correlation represented by the fields themselves, is a challenging task. Additionally, we are exploring ways to combine our method with machine learning strategies so that users will be able to identify and classify the correlations between multiple variables and the important data intervals for these variables.

Acknowledgements

This work was supported by the Lawrence Berkeley and Lawrence Livermore National Laboratories, and by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 through the Scientific Discovery through Advanced Computing (SciDAC) program's Visualization and Analytics Center for Enabling Technologies (VACET). We would like to thank colleagues in the Institute for Data Analysis and Visualization (IDAV) at UC Davis for their support during the course of this work.

References

[1] H. Akiba, K.-L. Ma, J. H. Chen, and E. R. Hawkes. Visualizing multivariate volume data from turbulent combustion simulations. *Computing in Science and Engineering*, 9(2):76–83, 2007.

[2] J. Becla and D. L. Wang. Lessons learned from managing a petabyte. In *CIDR*, pages 70–83, 2005.

[3] J. B. Bell, M. S. Day, I. G. Shepherd, M. R. Johnson, R. K. Cheng, J. F. Grcar, V. E. Beckner, and M. J. Lijewski. Numerical simulation of a laboratory-scale turbulent v-flame. *Proceedings of the National Academy of Science*, 102:10006–10011, July 2005.

[4] M. J. Berger and P. Colella. Local adaptive mesh refinement for shock hydrodynamics. *Journal of Computational Physics*, 82(1):64–84, May 1989.

[5] N. J. Brown, G. Li, and M. L. Koszykowski. Mechanism reduction via principal component analysis. *International Journal of Chemical Kinetics*, 29:393–414, 1997.

[6] M. Day and J. Bell. Simulation of premixed turbulent flames. *Journal of Physics Conference Series*, 46:43–47, Sept. 2006.

[7] H. Edelsbrunner, J. Harer, V. Natarajan, and V. Pascucci. Local and global comparison of continuous functions. In *IEEE Visualization*, pages 275–280. IEEE Computer Society, 2004.

[8] M. Frenklach, H. Wang, M. Goldenberg, G. P. Smith, C. Yu, D. M. Golden, C. T. Bowman, R. K. Hanson, D. Davidson, E. Chang, G. Smith, D. Golden, W. C. Gardiner, and V. Lissianski. Gri-mechan optimized detailed chemical reaction mechanism for methane combustion. Technical Report No. GRI-95/0058, Gas Research Institute, Nov. 1995.

[9] Y.-H. Fua, M. O. Ward, and E. A. Rundensteiner. Hierarchical parallel coordinates for exploration of large datasets. In *IEEE Visualization*, pages 43–50, 1999.

[10] L. Gosink, J. Shalf, K. Stockinger, K. Wu, and W. Bethel. Hdf5-fastquery: Accelerating complex queries on hdf datasets using fast bitmap indices. In *SSDBM*, pages 149–158, 2006.

[11] J. Gray, D. T. Liu, M. A. Nieto-Santisteban, A. S. Szalay, D. J. DeWitt, and G. Heber. Scientific data management in the coming decade. *SIGMOD Record*, 34(4):34–41, 2005.

[12] G. Hermosillo, C. Chefd'hotel, and O. Faugeras. Variational methods for multimodal image matching. *International Journal of Computer Vision*, 50(3):329–343, 2002.

[13] A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(2):69–91, 1985.

[14] N. Sauber, H. Theisel, and H.-P. Seidel. Multifield-graphs: An approach to visualizing correlations in multifield scalar data. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):917–924, 2006.

[15] K. Stockinger, E. W. Bethel, S. Campbell, E. Dart, and K. Wu. Imaging and visual analysis - detecting distributed scans using high-performance query-driven visualization. In *Supercomputing*, page 82. ACM Press, 2006.

[16] K. Stockinger, J. Shalf, E. W. Bethel, and K. Wu. Dex: Increasing the capability of scientific data analysis pipelines by using efficient bitmap indices to accelerate scientific visualization. In *Scientific and Statistical Database Management*, pages 35–44, 2005.

[17] K. Stockinger, J. Shalf, K. Wu, and E. W. Bethel. Query-driven visualization of large data sets. In *IEEE Visualization*, pages 167–174, 2005.

[18] R. Taylor. Visualizing multiple fields on the same surface. *IEEE Computer Graphics and Applications*, 22(3):6–10, 2002.

[19] T. Urness, V. Interrante, I. Marusic, E. Longmire, and B. Ganapathisubramani. Effectively visualizing multi-valued flow data using color and texture. In *IEEE Visualization*, pages 115–121, 2003.

[20] P. C. Wong and R. D. Bergeron. 30 years of multidimensional multivariate visualization. In G. M. Nielson, H. Hagen, and H. Müller, editors, *Scientific Visualization*, pages 3–33. IEEE Computer Society, 1994.

[21] K. Wu, W. S. Koegler, J. Chen, and A. Shoshani. Using bitmap index for interactive exploration of large datasets. In *Scientific and Statistical Database Management*, pages 65–74. IEEE Computer Society, 2003.

[22] K. Wu, E. J. Otoo, and A. Shoshani. On the performance of bitmap indices for high cardinality attributes. In M. A. Nascimento, M. T. Özsu, D. Kossmann, R. J. Miller, J. A. Blakeley, and K. B. Schiefer, editors, *Proceedings of the Thirtieth International Conference on Very Large Data Bases*, pages 24–35. Morgan Kaufmann, 2004.