

**DETC2011-48333**

**LIVE: A WORK-CENTERED APPROACH TO SUPPORT VISUAL ANALYTICS OF  
MULTI-DIMENSIONAL ENGINEERING DESIGN DATA WITH INTERACTIVE  
VISUALIZATION AND DATA-MINING**

**Xin Yan**

College of Information Sciences  
and Technology

**Mu Qiao**

Department of Computer Science  
and Engineering

**Timothy W. Simpson**

Department of Mechanical and  
Nuclear Engineering

**Jia Li**

Department of Statistics

**Xiaolong (Luke) Zhang**

College of Information Sciences  
and Technology

The Pennsylvania State University, University Park, PA, 16802, USA

**ABSTRACT**

*During the process of trade space exploration, information overload has become a notable problem. To find the best design, designers need more efficient tools to analyze the data, explore possible hidden patterns, and identify preferable solutions. When dealing with large-scale, multi-dimensional, continuous data sets (e.g., design alternatives and potential solutions), designers can be easily overwhelmed by the volume and complexity of the data. Traditional information visualization tools have some limits to support the analysis and knowledge exploration of such data, largely because they usually emphasize the visual presentation of and user interaction with data sets, and lack the capacity to identify hidden data patterns that are critical to in-depth analysis. There is a need for the integration of user-centered visualization designs and data-oriented data analysis algorithms in support of complex data analysis. In this paper, we present a work-centered approach to support visual analytics of multi-dimensional engineering design data by combining visualization, user interaction, and computational algorithms. We describe a system, Learning-based Interactive Visualization for Engineering design (LIVE), that allows designer to interactively examine large design input data and performance output data analysis simultaneously through visualization. We expect that our approach can help designers analyze complex design data more efficiently and effectively. We report our preliminary evaluation on the use of our system in analyzing a design problem related to aircraft wing sizing.*

**INTRODUCTION**

The "design by shopping" paradigm was first introduced by Balling [1], where a posteriori articulation of preference [2] is enabled to solve multi-objective optimization problems. This paradigm allows decision makers to view the visualization of the entire design space before choosing an optimal design based on their design preference. Comparing to traditional priori methods, the posteriori approach has the advantage of giving decision makers more control in the design exploration and selection process. However, the following problem in engineering design domain is that designers are now facing much more design alternatives simulated by fast computers than ever before, thanks to the rapid growth of computing power. The design and result information can easily overwhelm designers and go beyond human's analyzing capability. Selecting an optimal design candidate from such large scale design data sets becomes difficult for designers. Therefore, effective management and mining of massive data is needed in the process of design optimization.

To support the trade space exploration process, previous efforts have been made by applying multi-dimensional data visualization techniques [3,4]. Effective information visualization techniques are useful for designers' knowledge discovery and decision making tasks, mainly because human's visual perception capabilities are utilized to help identify patterns, trends, and features within datasets. However, due to the limitations in human cognition and displaying infrastructure, existing visualization systems lack the in-depth analyzing capability to discover knowledge from large-scale, unstructured data sets. Traditional approaches of multi-dimensional data visualization largely focus on visualizing

overall relationships among multiple dimensions, but often provide little support for quantitative examining the statistical characteristics of data distribution. Engineering design problems often involve multi-dimensional, continuous data, but powerful visualization tools to support such data sets are rare.

One possibly way to address such difficulty is to enhance information visualization by integrating fast computational methods, such as data mining algorithms. To extract useful information from large data sets, computational data mining methods have the advantage over visualization on capturing inherent patterns from big volume of data quickly. In the process of knowledge structuring and analytical reasoning, automatic techniques can help designers go beyond simple pattern detection and build high level hypotheses and complex models [5]. For instances, clustering can provide a grouping model where data is aggregated, and predicting rules can provide an inductive model where data is linked by if-then associations. These models are especially useful in engineering design problems. When designers are dealing with large amount of design alternatives, clustering can efficiently reduce the cognitive cost of examining them individually. Also, to find a specific group of good designs, designer needs to gradually narrow down to a relatively small subset in design space. Predicting rules like "what combination of data values can lead to target performance" can be useful guidance in the process of exploration and re-sampling design alternatives.

At the same time, integration of automatic data mining into interactive visualization faces some challenges. Because data-oriented computational algorithms and user-centered design interaction emphasize on different levels of the problem, there may be confliction between the two. Computational tools focus on the abstract, mathematical characteristics of datasets, such as data distributions, while interactive visualization designs concentrate on human users and the semantic meanings of data. Pure data mining methods lack the motivation to utilize ability of human brain to synthesize new knowledge [6]. Algorithms are often black-box to end users as well. As a result, the relevant expert knowledge of designers will be omitted. To avoid such problems, designer's control during the decision making process must be guaranteed. Another problem is that there may be cognitive difficulties for analysts to manipulate algorithms and understand results of them. Therefore, we need interactive visualization techniques to allow human designers intervening in the process of computational analysis and interpreting the results more easily.

To summarize, we need an efficient tool that integrates information visualization and data mining to support knowledge discovery and decision making in the context of engineering design. In this paper, we introduce a work-centered approach to support visual analytics of multi-dimensional engineering design data by combining visualization, user interaction, and computational algorithms. We also present a system prototype developed under this approach, Learning-based Interactive Visualization for Engineering design (LIVE), which allows designers to interactively explore the trade space

on different views with on-demand data clustering, classification, and aggregation. We also report our preliminary evaluation on the use of our LIVE system in analyzing a design problem related to aircraft wing sizing.

## RELATED WORK

Related research concerns the areas of visualization in engineering design, visual data mining and visual analytics, and current data mining and clustering algorithms.

### Visualization in Engineering Design

In previous research, visualization has been used to support posteriori methods of multi-objective optimization in engineering design tasks. For instances, the ATSV system [3,4] is developed to visualize multi-dimensional trade spaces by combination of histogram, scatter plots, parallel coordinate plots, etc; Hyper-space Diagonal Counting (HSDC) method [7] is proposed to improve the visualization of a Pareto Frontier for an  $n$ -dimensional performance space. Visualization is useful in multi-objective optimization processes, because efficient information visualization amplifies human cognition and facilitates user's understanding on features of data. However, to enable effectively support decision making in engineering design, information visualization must satisfy some formulated requirements [8], such as simplicity, persistence and completeness. As the scale of data sets increases, it becomes difficult for existing visualization techniques to satisfy such requirements. Current information visualization tools often lack the ability to deal with multi-dimensional, unorganized data sets.

More systematic frameworks have also been described in previous literatures. Visual steering commands [9,10] provide designers various user guided samplers to explore the trade space and exploit new information and insights. The framework of VIDEO [11] allows users to navigate multi-objective solution sets visually and to identify one or more optimal designs. These tools attempt to provide designers better control during the decision making process by constructing more interactive environment. A framework of interactive multi-scale visualization [12] is also presented to support the knowledge exploration and discovery in engineering design. However, these systems hardly have in-depth data analysis capacities.

One the other hand, some research tries to reduce data sets by simple aggregating. Cloud Visualization [13] and BrickViz [14] explore methods for grouping and aggregating data to reduce users' cognitive workload; however, neither approach is supported by formal clustering techniques. Chiu and his co-authors [15,16] provide a novel means for aggregating hyper-dimensional data and visualizing Pareto fronts that span  $n$ -dimensions and steering subsequent searches [17]. Meanwhile, Ross, et al. [18] have introduced a framework for multi-attribute trade space exploration. The emphasis in their work is on the use of multi-attribute utility theory to integrate designers' preferences for multiple objectives, not the visualization of the results, *per se*. In complex design

optimization problems, more advanced tools to integrate automatic data mining algorithms are still needed by designers seeking more comprehensive guidance such as summarizing data distributions, clustering, etc.

### **Visual Data Mining and Visual Analytics**

The idea of integrating information visualization and data mining was proposed by researchers from both communities recently [19,20]. This is based on the fact that human and computer are complementary in the process of knowledge discovery. By adding data mining algorithms into visualization, overwhelming information can be reduced so that patterns become more prominent in limited visual configuration. For instance, feature selection is used to hierarchically filter dimensions when exploring high dimensional data sets [21]; automatically detected clusters are displayed with colors when visualizing social networks [22]. The limitation of such works is their ad-hoc nature to a large extent. Systems supporting the entire knowledge discovery and decision making process are rare.

In the past ten years, a new research field of visual analytics [23] has emerged with a focus on analytical reasoning facilitated by interactive visual interfaces. The goal of visual analytics aims to integrate disciplines such as visualization, data management, and data mining to solve problems of complex reasoning and decision making. While still in its infancy, current visual analytics faces various difficulties. Particularly, supporting interactive visualization of large databases is inadequately addressed [24]. Such a system requires not only fast feedbacks during user's exploration, but also meaningful solutions to scale down both data cardinality and data dimensions.

### **Statistical Clustering**

One of often used computational data mining algorithms in data analysis is statistical clustering. Clustering, as one of the most fundamental techniques in data mining, has been employed in tremendously diverse areas for a multitude of purposes. Clustering groups a set of objects into subsets (aka clusters) so that the objects in the same subset share similar characteristics. It reveals the essential information or underlying patterns from massive data so that the subsequent analyses or processes become feasible or efficient. For instance, clustering can speed up images indexing and retrieval [25]. Clustering can also serve as stand-alone processes. For instance, text documents are often clustered based on their underlying topics [26]. The methodologies, applications, and practices in clustering can be found in recent surveys [27-29].

Generally speaking, there are three types of clustering methods. The first type focuses on optimizing a given objective function, which often reflects the general criteria about a good clustering, i.e., the objects in the same cluster should be similar while those in the different clusters be as distinct as possible. K-means and k-center [30] are representatives in this type. The second type only uses the pairwise distances between objects to

do the clustering. These methods have enjoyed wide applications when a mathematical representation for the object is difficult or intractable to obtain. Examples are Linkage clustering [31] and spectral graph partitioning [32]. However, these methods do not scale well with large data sets due to the quadratic computational complexity of calculating the pairwise distances. The third type is based on statistical modeling. The probability density function of the entire data is modeled as a mixture of several parametric distributions [33], e.g., mixture of Gaussians for continuous data and mixture of Poissons for discrete data. The clustering procedures involve first fitting a mixture model with the expectation-maximization (EM) algorithm and then computing the posterior probability of each mixture component given the data point. The component with the highest posterior probability is chosen for that data point and all the data points belonging to the same component form one cluster. Moreover, the posterior probability of each component can be regarded as a soft clustering scheme. In addition to clustering data, a probability distribution is obtained for each cluster, which can be useful to gain some insight about the data. The clustering algorithm we have employed in this research, mode association clustering (MAC), belongs to the third type of clustering, which is a new nonparametric statistical approach [34].

In the context of dealing with large-scale design data sets, clustering may efficiently reduce designer's cognitive load by grouping similar design alternatives together. Designers can also uncover hidden patterns between design input and performance output variables when viewing a cluster of similar designs at the same time. Nonetheless, to satisfy ad-hoc design goals, a human designer's intervention needs to be introduced into the computation of these automatic clustering algorithms. Results of algorithms are also expected to be easily interpretable.

In summary, current approaches are often weak to support effective and efficient knowledge discovery when exploring massive engineering design data sets. More advanced tools aiming at such complex tasks are therefore needed.

## **WORK-CENTERED APPROACH**

We will pursue a work-centered approach to support trade space exploration by integrating data-oriented computational methods with interactive visualization. In this section, we first briefly present our work-centered model, and then discuss the implementation of the LIVE system guided by the model.

### **Work-Centered Model**

Figure 1 illustrates our work-centered model. Compared with traditional visualization models, e.g., the reference model by Card et al. [35], our work-centered model emphasizes the integration of data-oriented computational algorithms (e.g., data-mining, data-clustering) with user-centered designs. The data-oriented analysis algorithms can identify useful patterns hidden in massive data, while the user-centered designs are

concerned with the perceptual, cognitive, and task characteristics of human users in computing systems. Through this integration, we aim at maximizing the strengths of both human judgment and computing power.

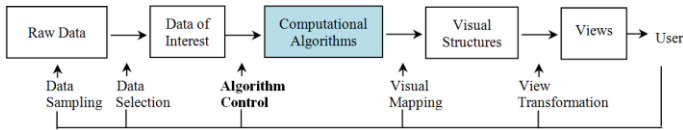


Figure 1. THE FRAMEWORK FOR WORK-CENTERED APPROACH TO SUPPORT TRADE SPACE EXPLORATION

## LIVE SYSTEM

Based on this model, we developed the LIVE system. Our design is based on the observations on typical engineering design processes. General trade space exploration consists of three main steps. First, a simulation model that carries unknown relationship between design input variables and performance output variables is constructed. Then, experiments are run to generate large amount of design alternatives with their corresponding performance parameters. At last, the designer explores the trade space using interactive visualization tools to find the patterns of design data and data dimensions and to identify preferred designs.

The LIVE system focuses on supporting two key tasks in trade space exploration: design input analysis and performance output analysis.

- **Design Input Analysis.** With large amount of design alternatives displayed, a designer often wants to aggregate designs into groups and see whether designs from the same group can get similar performances.
- **Performance Output Analysis.** Designers also want to specify a group of interested designs in performance space according to their preferences. Then they are interested in knowing what combinations of design values may lead to similar performance.

Central of supporting these two tasks are a decision-tree-based data classification algorithm and a mixture model and mode association based clustering algorithm. Generally speaking, a decision-tree algorithm uses a tree-style model to show the possible decision consequences based on given conditions. Under our scenario, this method can be used to analyze how design inputs may affect design outcomes. To help designers see what the possible design results are, a decision tree can classify the input variables based on their values and show the design outputs available from each class of design inputs. The mixture-model and mode association based clustering algorithm can identify possible data clusters without the need to distinguish inputs and outputs. This method can be used to analyze the distribution patterns of design outputs and help designers see potential relationships.

## Decision-Tree Classification

The decision tree is a tree-like structure. Each node in it corresponds to an input attribute with a splitting value. Each leaf of the tree specifies the expected output value, as a consequence of the particular input values described by the path from the root to that leaf. In our case, a designer selects on values of output performance variables, and the decision tree divides the design space into hierarchical regions. In this way, the classifier generated from decision-tree algorithm can be interpreted easily. Every node is described by a hyperrectangle in the multidimensional space. The hierarchical structure is also useful for the understanding and analytical process. At each leaf node, an estimate for the posterior probability of classes is calculated at the same time. Hence, users can see how pure each node is.

In LIVE, we integrated C4.5 algorithm [36] to build a decision tree upon the designer's selection of a group of interested designs. The basic idea behind C4.5 algorithm is that in the target decision tree, each node is associated with an input attribute that is *most informative* among the attributes for data traced into this node. The amount of information is measured by Shannon *entropy*, firstly introduced by Claude Shannon in Information Theory. For a random variable  $X$  with  $n$  values  $\{x_i : i=1, \dots, n\}$ , the Shannon entropy  $H(X)$  is defined as

$$H(X) = -\sum_{i=1}^n p(x_i) \log_b p(x_i) \quad (1)$$

where  $p(x_i)$  is the probability mass function at  $x_i$ .

The algorithm then approaches in a greedy way to iteratively search all of the attributes in the training data set and choose one that splits the current node of data most effectively. The criteria for splitting the data is to select the attribute that results in the largest *information gain*,  $IG(Y|X)$ , defined as the difference in information entropy.

$$IG(Y|X) = H(Y) - H(Y|X) \quad (2)$$

After setting one split attribute, the system recursively searches for all subspaces and arrange the rules in a tree structure. The process will only stop when some condition is reached, such as the number of instances in a subspace is less than a threshold, or the classification error rate is lower than a confidence threshold.

## Mode Association Clustering Algorithm

Our work here focused on the implementation and integration of a hierarchical mode association clustering (HMAC) algorithm. Developed in [34], the HMAC algorithm is a new mixture model based, nonparametric statistical clustering method. Comparing with the conventional approach, the mixture components serve the sole purpose of accurately estimating the overall density instead of attempting to characterize each cluster simultaneously. We have developed a new optimization algorithm to find the local modes of mixture models and to find the ridgeline linking two hilltops. Data points are grouped into one cluster if they are associated with

the same mode. This framework enables more precise representation of clusters and refined modeling of the entire data. The distributions of individual clusters are no longer limited to certain assumed forms, e.g., Gaussian. In addition, the ridgeline, in its own right, is an effective visualization tool to demonstrate the geometry of high dimensional probability distributions, enabling diagnosis of the separateness between different clusters.

Given a data set  $\{x_1, x_2, \dots, x_n\}$ ,  $x_i \in R^d$ , a probability density function for the data is estimated non-parametrically using Gaussian kernels. As the kernel density estimate is in the form of a mixture distribution, an iterative process can be initiated to find a mode using every sample point  $x_i$ ,  $i = 1, 2, \dots, n$ . Points sharing the same mode will be grouped into one cluster. When the variances of Gaussian kernels increase, the density estimate becomes smoother and tends to group more points into one cluster. A hierarchy of clusters can thus be constructed by gradually increasing the variances of Gaussian kernels.

Assume the set of data to be clustered such that  $S = \{x_1, x_2, \dots, x_n\}$ ,  $x_i \in R^d$ . The Gaussian kernel density estimate is formed:

$$f(x) = \sum_{i=1}^n \frac{1}{n} \phi(x | x_i, \Sigma) \quad (3)$$

where the Gaussian density function is:

$$\phi(x | x_i, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - x_i)' \Sigma^{-1}(x - x_i)\right) \quad (4)$$

Here,  $\Sigma$  is a spherical covariance matrix  $\text{diag}(\sigma^2, \sigma^2, \dots, \sigma^2)$  and  $\sigma$  is the standard deviation, also referred to as the *bandwidth* of the Gaussian kernel.

When the bandwidth  $\sigma$  increases, the kernel density estimate becomes smoother and more points tend to climb to the same mode. This suggests a natural approach for hierarchical clustering. Given a sequence of bandwidths  $\sigma_1 < \sigma_2 < \dots < \sigma_n$ , hierarchical clustering is performed in a bottom-up manner. Such hierarchical clustering allows multi-scale data clustering.

Compared with other clustering methods, this mixture modeling approach enjoys broader applications because this clustering method can be used for both unsupervised and supervised data sets. Even for data clustering, the mixture components can be used as probabilistic descriptions of data clusters, which provide users with deeper insight into the characteristics of data clusters.

Another benefit of this mixture modeling method is that it allows users to weight data dimensions differently. By choosing different weights for individual dimension, the designer can examine how each dimension may influence clustering results differently.

### Interactive Visualization

The goal of visualization interface is to provide designer quick and understandable ways to access data and algorithms.

There are several key issues we emphasized on the design of visualization.

**Data Presentation.** LIVE system presents the entire trade space for the designer's exploration using multi-dimensional visualization techniques. The dimensions of engineering design data can usually be separated into two parts, the design input parameters, and the performance output parameters. Thus, it would be efficient for designers to display these two data spaces separately. The designer should be able to change the view of data from different aspects.

**Data Linking.** Designer not only needs a clear view of both design space and performance space, but also wants to establish explicit links between the two. These links in the designer's mind are often structured with help from effective visualization tools. Using the same visual glyphs is a practical way to connect the data in different spaces. When the system quickly maps designer's selection from one space to the other, the designer is able to recognize what design values may lead to aimed performances. Specifically, when designer acts on one of the plots, coincident points are highlighted on the other plot immediately. In-depth analysis is also performed in two spaces correspondingly. For instance, selecting a leaf node in the decision tree, data points in that subspace are highlighted in both plots using the same way of representation; when a group of clustered design alternatives is selected, corresponding points in performance space are highlighted instantly.

**Data Filtering.** Data filtering is an important way to help designer concentrate on a specific set of interested designs. Simple filtering can be carried out by setting numerical constraints through manipulating several slider bars associated with design and performance variables. Designer can visually pop out the interested group and fade out unimportant data. To be more flexible, data mining algorithms can be applied to a particular subset as well.

**User's Control on Algorithms.** The integration of data mining algorithms and visualization is mainly by means of providing user various ways to control the algorithms and refine his hypotheses. Through this way the designer is able to add his preference and domain knowledge into the calculation of data mining. In our system design, putting human in the loop means the designer is able to initial the algorithm, specify inputs, adjust parameters, and see real-time results. In LIVE, the designer determines when and how to perform the computation such as decision tree constructing and mode association clustering. For example, in the process of building decision tree, each selection of the designer refreshes the two-class labels of the training data: interesting versus non-interesting design. The designer can decide which dimensions to include or exclude in the calculation of the decision tree. The designer can also decide whether to prune the tree at run time. In clustering calculation, the designer can perform the automatic clustering and dynamically specify the bandwidth used in mode association. We use TreeMap [37] to visualize the



Figure 2. SCREENSHOT OF LIVE SYSTEM: (a) SCATTER PLOT OF PERFORMANCE OUTPUT VARIABLES; (b) SCATTER PLOT OF DESIGN INPUT VARIABLES (SHOWING CLUSTERING RESULTS); (c) TEXT REPRESENTATION OF DECISION RULES; (d) TREEMAP VISUALIZATION OF DECISION TREE; (e) ADJUSTABLE SLIDERS OF RANGES OF INPUT AND OUTPUT VARIABLES.

decision tree and color coding to visualize the clustering results. The designer can view and manipulate area of interest by interacting with the visual components in these visualizations as well.

## USER INTERFACE DESIGN

Based on the considerations, we designed the user interface of LIVE, shown in Figure 2. The whole user interface is vertically divided into two main parts. On the top are two scatter plots displaying data points of performance output space (Panel a) and design input space (Panel b). On the bottom are the visualization tools to present decision tree analysis results (Panel c and d) and to control clustering parameters (Panel e).

### Interactive Scatter Plots and Scatter Plot Matrix

The data sets from design and performance spaces are firstly visualized in two scatter plots on the top of the interface window (Figure 2a, Figure 2b). All data points in the space are shown as gray dots. Different visual glyphs, such as green cross, blue diamond, and red box, are used to highlight designer's selection or statistic analysis results. The designer can specify which two variables are shown as the X and Y axis at any time. The scatter plot matrix of all input/output dimensions is also available on user's diamond (Figure 3) to give designer a full view of data from different perspectives. Both of

the input and output scatter plots are interactive. The designer can select rectangle or polygon areas of interest in either plot. Subplots in the scatter plot matrix are also selectable. Selecting on scatter plot matrix is more flexible than on two dimension plots. Selection on scatter plot matrix can be directly added to the main plot for further analysis.

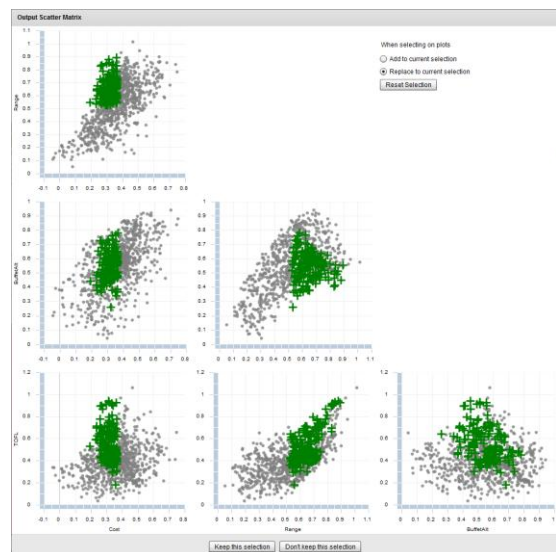


Figure 3. INTERACTIVE SCATTER PLOT MATRIX

## Data Selecting and Data Linking

The designer can select a rectangle or polygon region on output and input plot by mouse dragging. He can also specify a group of designs by narrowing down the output or input sliders (Figure 2e). Three types of glyphs are used to distinguish designer's different selection actions. The designer selected points on the performance space (right plot) are displayed with green crosses. Selections on the design space (left plot) are highlighted using red boxes. Finally, the designer's selection on slider bars and on decision tree nodes are shown by blue diamonds. To help a designer maintain his mental model of data between different plots, as long as the designer selects on one plot, the corresponding data points are mapped and highlighted in the other plot with the same glyph immediately. Also, after the designer performs the clustering calculation, he can specify a group of aggregated design alternatives by double clicking on any data points in the cluster. The selected group of designs is then highlighted in the performance space. Figure 5 is a sample screenshot of data mapping, which shows designer's selection on output (green crosses) and input plots (red dots).

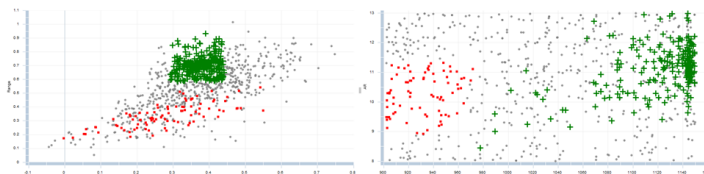


Figure 5. SCREEN SHOT OF LINKING DESIGNER'S SELECTION ON DIFFERENT PLOTS

## Visualization of Decision Tree

The decision tree is constructed at run time to provide designers rules of guidance for predicting what combination of design values may lead to performance output in a user specified range. The interactive process begins with designer's selecting on output plot. Either mouse dragging direct on output scatter plot or moving sliders (Figure 6) can be used to select an interested performance group as input to the algorithm. The decision tree provides a segmentation of the whole data space. Basically, each interior node in the decision tree corresponds to a split in one input variables. Each leaf node in the tree represents a combination of rules for classifying the input variables. The rules are extracted from the decision associated with each node along the path from root to leaf. The leaf nodes also include information about how many data samples following the rules will reach the target output values. During the calculation, designer can also specify whether to prune the tree by clicking the checkbox.



Figure 6. SPECIFYING PERFORMANCE GROUPS USING SLIDERS

The result tree is visualized in a format of TreeMap [37], which is a visualization method of displaying hierarchical structured data using nested rectangles. In our interface, each leaf node in a decision tree is visualized as a rectangle in TreeMap. The size of the rectangle represents the number of design samples following the input classifying rules. The more samples fall into the node, the bigger the node. The color of the rectangle represents the ratio between number of samples falling into user specified performance over the total number in this node. Leaf nodes with high percentage of hit are shown in green, and low hit nodes are in red. Underneath the TreeMap there is also a legend bar showing the relationship between color and hit ratio. Figure 6a and Figure 6b display a high hit leaf node (97.29%) and a low hit leaf (2.70%), respectively.

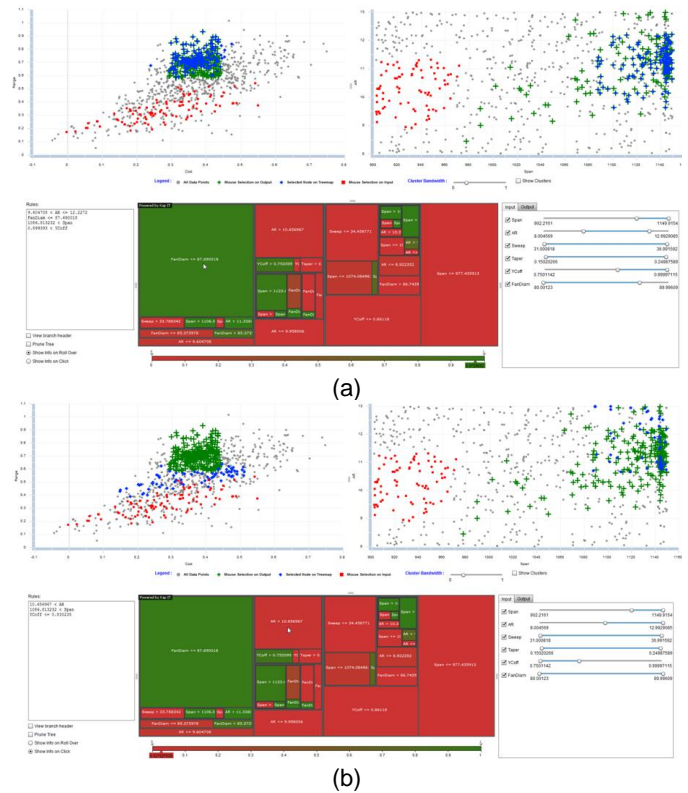


Figure 4. DIFFERENT TYPES OF NODE IN A DECISION TREE: (a) A HIGH HIT LEAF NODE, (b) A LOW HIT LEAF NODE

The entire TreeMap is interactive, which means designer can click any node on it. After the clicking, all design samples in this leaf node will be highlighted as blue diamonds in both input and output plots. The rules of input variables are shown in plain texts to the left of TreeMap, and also displayed as slider bars which are to the right. There are checkboxes in front of each slider bar to allow designer include or exclude any variables in the calculation of decision tree. When change is made, the decision tree will be re-constructed. Furthermore, the designer can adjust the sliders bars to refine area of interest. Data highlighting in plots changes consequently.

## Visualization of Data Clustering

In previous work we have developed a nonparametric clustering method by mode identification [34]. The clustering algorithm groups data points into one cluster if they are associated with the same local maximum of the kernel density estimation function. This method can be extended for hierarchical clustering by gradually increasing the bandwidths of kernel density estimators. We apply this Hierarchical Mode Association Clustering (HMAC) approach to aggregate design data space. The designer can view the clustered data set by simply checking the "show clusters" box under input plot. He can also change the bandwidth to get different levels of clustering aggregation. Figure 8 shows two clustering results, with two bandwidths, of a wing design data sets with six data dimensions. Here we only map two data dimensions into a scatter plot and color-code the data clusters produced by the algorithm. Figure 8 shows how different data clusters are distributed. By increasing the bandwidth of clustering methods, users are able to get a result with fewer clusters and see different data distribution patterns. Clustering can also be performed only on selected designs, which allows designer to concentrate on analyzing those interested design alternatives only by hiding unselected data.

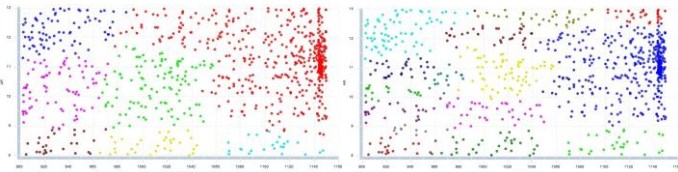


Figure 8. MULTI-SCALE DATA CLUSTERING RESULTS OF MULTI-DIMENSIONAL DATA WITH A MIXTURE MODEL ALGORITHM

## System Implementation

The whole system is implemented in a client-server model. The client part mainly includes interactive visualization interface, while the server side manages data storage and computing jobs. The client sends user actions to the server and gets computing results back through web communications. The entire interface is web-based, platform-independent so that designers can access it with any platform. The computation tasks such as clustering and decision tree are completed remotely on the server side. Algorithms are implemented in Java. We borrowed part of the source code of J48 in Weka data mining package [38] to implement C4.5 algorithm for decision tree building.

## CASE STUDY

To understand how the LIVE system can support the knowledge discovery and decision making, we conducted a preliminary study by using a practical engineering design problem of aircraft wing sizing. This problem [39] involves sizing an aircraft wing's plan view layout. There are six design variables as inputs, which are shown in Figure 8.

1. Semi-span:  $900 \leq Span \leq 1150$
2. Aspect ratio:  $8 \leq AR \leq 13$
3. Quarter chord sweep angle:  $31 \leq Sweep \leq 37$  (5)
4. Taper ratio:  $0.15 \leq Taper \leq 0.25$
5. Sparbox root chord:  $0.75 \leq YCoff \leq 1$
6. Fan diameter:  $80 \leq FanDiam \leq 90$

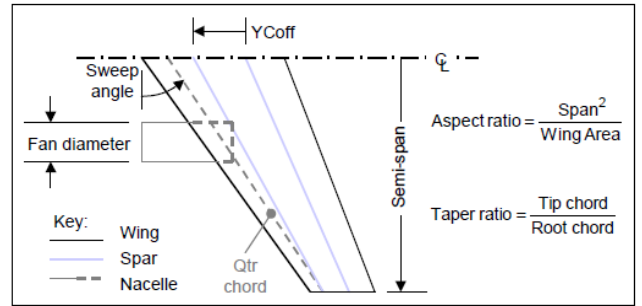


Figure 7. DESIGN INPUT VARIABLES IN WING SIZING PROBLEM

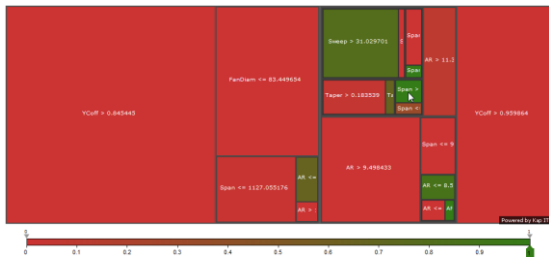
Four output performance variables of the problem are *Cost*, *Range*, *Buffet altitude* and *Takeoff field length (TOFL)*. The general multi-objective problem is stated as [40]:

$$\begin{aligned}
 \text{Goal:} & \quad \text{Minimize } Cost \text{ while Maximize } Range \\
 \text{Subject to:} & \quad Range \geq 0.589 \\
 & \quad Buffet \text{ altitude} \geq 0.603 \\
 & \quad Takeoff \text{ field length} \leq 0.377
 \end{aligned}
 \tag{6}$$

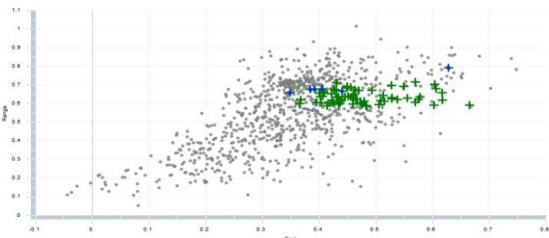
Let us assume the following usage scenario. In our system, after importing the computer simulated data set of design alternatives, the designer can at first select feasible designs among the entire design space by adding three numerical constrains (Figure 6).

A decision tree is then calculated and visualized (Figure 9a) to guide a designer so that he can see which combinations of design variables can lead to performances within feasible designs. Based on the result, the designer can compare several suggested high hit areas shown in green. Reminded of the design goal of minimizing *Cost* while maximizing *Range*, the designer picks an area including 6 designs, where the mouse is pointing to in Figure 9a. The design alternatives are highlighted in blue in Figure 9b. The corresponding rules on design variables are:

$$\begin{aligned}
 AR & \leq 11.339, \\
 FanDiam & > 86.019, \\
 Span & > 1132.837, \\
 Sweep & \leq 32.089, \\
 0.911 & < YCoff \leq 0.960.
 \end{aligned}
 \tag{7}$$



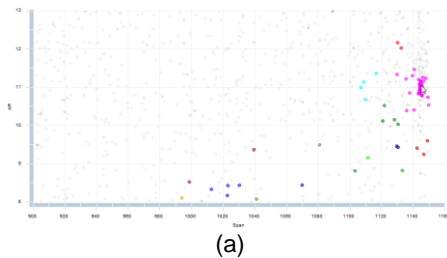
(a)



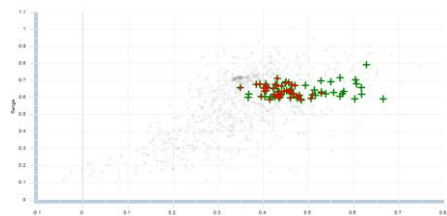
(b)

Figure 9. (a) DECISION TREE RESULT; (b) A GROUP OF DESIGNS MAPPING ON PERFORMANCE SPACE

Another way of exploring is by viewing the clustering on feasible designs in input space. As shown in Figure 10, the infeasible designs are faded out and the system automatically clusters feasible designs (Figure 10a). To fit the design goal, the designer can also examine clusters for an interested group of designs and their reflection on output performance space (Figure 10b).



(a)



(b)

Figure 10. (a) FEASIBLE DESIGN CLUSTERING RESULT; (b) CLUSTERED DESIGNS MAPPING TO PERFORMANCE SPACE

## CONCLUSION AND FUTURE WORK

This paper described a work-centered visual analytics approach on engineering data. We developed a system prototype, LIVE, to support knowledge discovery and decision making in engineering design. With our system, designers can follow a "design by shopping" paradigm to identify interested

solutions by exploring the entire trade space without priori specification of preferences, and the LIVE system attempts to address the problem of information overload in this process. To help designers make sense of large-scale, multi-dimensional engineering design data sets, the LIVE system tries to integrate user-centered interactive visualization with data-oriented computational algorithms.

This research has the potential to advance the visual analytics of the dependencies of continuous data in engineering, and foster the exploration and discovery of new knowledge and solutions for various engineering design problems. Our model and system will offer new approaches to gain important insights into the hidden relationships that influence the performance of engineered systems. Such insights can greatly benefit engineering design and allow the reduction of decision space to explore, the acceleration of decision-making, the avoidance of uncertainty, and the identification of key investable resources. By integrating user-centered design and data-oriented data-processing algorithms, this work will reconcile human users' limited capacity to process large amount and rapid growth of information in decision making, and lead to a new paradigm in visualization design that potentially maximizes the strengths of both human judgment and computing power.

Future work of this research can be conducted in several directions. Firstly, a more complete validation study of our design is necessary. In this paper we demonstrate how our system works with a real design problem of aircraft wing sizing. The new system is supposed to facilitate designer's tasks of trade space exploration. Further evaluation can be carried out to see how our system supports the discovery of hidden data features and enhances designer's performance through focus group observation and field study. Secondly, a full mixed-initiative integration of visualization and data mining [41] can be pursued. The LIVE system is an attempt to introduce data mining algorithms into traditional visualization supported design paradigm. It can be improved by allowing users to become more involved in the process of data model constructing, so that users have more flexibility to fit different design requirements.

## ACKNOWLEDGMENTS

The authors acknowledge support by a grant from National Science Foundation (CCF 0936948).

## REFERENCES

- [1] Balling, R., 1999, "Design by Shopping: A New Paradigm?", *Proceedings of the Third World Congress of Structural and Multidisciplinary Optimization (WCSMO-3)*, Buffalo, NY, University at Buffalo, 295-297.
- [2] Hwang, C.-L. and Masud, A. S., 1979, *Multiple Objective Decision Making - Methods and Applications*, New York, Springer-Verlag.
- [3] Stump, G., Yukish, M. and Simpson, T. W., 2004, "The ARL Trade Space Visualizer: An Engineering Decision-Making Tool", *10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, Albany, NY, AIAA, AIAA-2004-4568.

- [4] Stump, G. M., Yukish, M., Simpson, T. W. and Bennett, L., 2002, "Multidimensional Visualization and Its Application to a Design by Shopping Paradigm", *9th AIAA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, Atlanta, GA, AIAA, AIAA-2002-5622.
- [5] Bertini, E. and Lalanne, D., 2010, "Investigating and reflecting on the integration of automatic data analysis and visualization in knowledge discovery", *ACM SIGKDD Explorations Newsletter*, 11 (2), 9-18.
- [6] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., 1996, "The KDD process for extracting useful knowledge from volumes of data", *Magazine of Communications of the ACM*, 39 (11), 27-34.
- [7] Agrawal, G., Lewis, K., Chugh, K., Huang, C.-H., Parashar, S. and Bloebaum, C. L., 2004, "Intuitive Visualization of Pareto Frontier for Multi-Objective Optimization in n-Dimensional Performance Space", *10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, Albany, NY, AIAA, AIAA-2004-4434.
- [8] McQuaid, M.J., Ong, T.-H., Chen, H., Nunamaker, J.F., 1999, "Multidimensional scaling for group memory visualization", *Journal of Decision Support Systems*, 27, 163-176.
- [9] Stump, G., Lego, S., Yukish, M., Simpson, T. W. and Donndelinger, J. A., 2007, "Visual Steering Commands for Trade Space Exploration: User-Guided Sampling with Example", *ASME Design Engineering Technical Conferences - Design Automation Conference*, Las Vegas, NV, ASME, DETC2007/DAC-34684.
- [10] Simpson, T. W., Spencer, D. B. and Yukish, M. A., 2008, "Visual Steering Commands and Test Problems to Support Research in Trade Space Exploration", *12th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, Victoria, British Columbia, Canada, AIAA, AIAA-2008-6085.
- [11] Kollat, J. B. and Reed, P., 2007, "A framework for Visually Interactive Decision-making and Design using Evolutionary Multi-objective Optimization (VIDEO)", *Journal of Environmental Modelling & Software*, 22 (12) (December 2007), 1691-1704.
- [12] Zhang, X., Simpson, T. W., Frecker, M. and Lesieutre, G., 2010, "Supporting knowledge exploration and discovery in multi-dimensional data with interactive multiscale visualisation", *Journal of Engineering Design*.
- [13] Eddy, J. and Lewis, K., 2002, "Visualization of Multi-Dimensional Design and Optimization Data Using Cloud Visualization", *ASME Design Engineering Technical Conferences - Design Automation Conference*, Montreal, Quebec, Canada, ASME, September 29 - October 2, Paper No. DETC02/DAC-02006.
- [14] Kanukolanu, D., Lewis, K. E. and Winer, E. H., 2006, "A Multidimensional Visualization Interface to Aid in Trade-off Decisions During the Solution of Coupled Subsystems Under Uncertainty", *ASME Journal of Computing and Information Science in Engineering*, 6(3), 288-299.
- [15] Chiu, P.-W., Naim, A. M., Lewis, K. E. and Bloebaum, C. L., 2009, "The Hyper-Radial Visualization Method for Multi-Attributed Decision-Making under Uncertainty", *International Journal of Product Development*, 9(1-3), 4-31.
- [16] Chiu, P.-W. and Bloebaum, C. L., 2010, "Hyper-Radial Visualization (HRV) Method with Range-based Preferences for Multi-objective Decision Making", *Structural and Multidisciplinary Optimization*, 40(1-6), 97-115.
- [17] Chiu, P.-W. and Bloebaum, C. L., 2009, "Visual Steering for Design Generation in Multi-Objective Optimization Problems", *47th AIAA Aerospace Sciences Meeting*, Orlando, FL, AIAA, AIAA-2009-1167.
- [18] Ross, A. M., Hastings, D. E., Warmkessel, J. M. and Diller, N. P., 2004, "Multi-Attribute Tradespace Exploration as Front End for Effective Space System Design", *Journal of Spacecraft and Rockets*, 41(1), 20-28.
- [19] Shneiderman, B., 2001, "Inventing Discovery Tools: Combining Information Visualization with Data Mining", *Lecture Notes in Computer Science*, 2001, Volume 2226/2001, 17-28.
- [20] Ferreira de Oliveira, M.C. and Levkowitz, H., 2003, "From visual data exploration to visual data mining: a survey", *IEEE Transactions on Visualization and Computer Graphics*, 9 (3), 378-394, July-Sept.
- [21] Yang, J. and Ward, M. O. and Rundensteiner, E. A. and Huang, S., 2003, "Visual hierarchical dimension reduction for exploration of high dimensional datasets", *Proceedings of the symposium on Data visualisation*, 19-28, Eurographics Asso., Aire-la-Ville, Switzerland.
- [22] Heer, J. and Boyd, D., 2005, "Vizster: Visualizing Online Social Networks", *IEEE InfoVis 2005*, 5, 2005.
- [23] Thomas, J. J. and Cook, K. A. (editors), 2005, *Illuminating the path: The research and development agenda for visual analytics*, National Visualization and Analytics Center (NVAC).
- [24] Keim, D. A., Kohlhammer, J., Ellis, G. and Mansmann, F. (editors), 2010, *Mastering the Information Age - Solving Problems with Visual Analytics*, Eurographics.
- [25] Li, J., 2005, "Clustering based on a multi-layer mixture model", *Journal of Computational and Graphical Statistics*, 14(3), 547-568.
- [26] Blei, D. M., Ng, A. Y., Jordan, M. I., and Lafferty, J., 2003, "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, 3, 993-1022.
- [27] Kettnering, J. R., 2006, "The practice of cluster analysis", *Journal of Classification*, 23(1), 3-30.
- [28] Jain, A. K., Murty, M. N., and Flynn, P. J., 1999, "Data clustering: A review", *ACM Computing Surveys*, 31(3), 264-323.
- [29] Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P., 1996, *From data mining to knowledge discovery: An overview*, Advances in Knowledge Discovery, 1-35, AAAI/MIT Press.
- [30] Gonzalez, T. F., 1985, "Clustering to minimize the maximum intercluster distance", *Theoret. Comp. Sci.*, 38(22), 293-306.
- [31] Gower, J. C. and Ross, G. J. S., 1969, "Minimum spanning trees and single linkage cluster analysis", *Applied Statistics*, 18(1), 55-64.
- [32] Pothan, A., Simon, H. D., and Liou K., 1990, "Partitioning sparse matrices with eigenvectors of graphs", *SIAM Journal on Matrix Analysis and Applications*, 11(3), 430-452.
- [33] McLachlan, G. J. and Peel, D., 2000, *Finite Mixture Models*, New York, Wiley.
- [34] Li, J., Ray, S. and Lindsay, B., 2007, "A nonparametric statistical approach to clustering via mode identification", *Journal of Machine Learning Research*, 8(8), 1687-1723.
- [35] Card, S.K., Mackinlay, J. D., and Shneiderman, B., 1999, *Readings in information visualization: using vision to think*, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.
- [36] Quinlan, J. R., 1993, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- [37] Shneiderman, B., 1992, "Tree visualization with tree-maps: 2-d space-filling approach", *ACM Tran. on Graphics*, 11 (1), 92-99.
- [38] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H., 2009, "The WEKA Data Mining Software: An Update", *SIGKDD Explorations*, Volume 11, Issue 1.
- [39] Simpson, T. W. and Meckesheimer, M., 2004, "Evaluation of a Graphical Design Interface for Design Space Visualization", *45th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics & Materials Conference*, Palm Springs, CA, AIAA, AIAA-2004-1683.
- [40] Simpson, T. W., Carlsen, D. E., Congdon, C. D., Stump, G. and Yukish, M. A., 2008, "Trade Space Exploration of a Wing Design Problem Using Visual Steering and Multi-Dimensional Data Visualization", *4th AIAA Multidisciplinary Design Optimization Specialist Conference*, Schaumburg, IL, AIAA, AIAA-2008-2139.
- [41] Bertini, E. and Lalanne, D., 2010, "Investigating and reflecting on the integration of automatic data analysis and visualization in knowledge discovery", *SIGKDD Explorations*, 11(2), 9-18.