# Progressively Consolidating Historical Visual Explorations for New Discoveries

Kaiyu Zhao, Matthew O. Ward, Elke A. Rundensteiner, and Huong N. Higgins

Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA, USA

## ABSTRACT

A significant task within data mining is to identify data models of interest. While facilitating the exploration tasks, most visualization systems do not make use of all the data models that are generated during the exploration. In this paper, we introduce a system that allows the user to gain insights from the data space progressively by forming data models and consolidating the generated models on the fly. Each model can be a a computationally extracted or user-defined subset that contains a certain degree of interest and might lead to some discoveries. When the user generates more and more data models, the degree of interest of some portion of some models will either grow (indicating higher occurrence) or will fluctuate or decrease (corresponding to lower occurrence). Our system maintains a collection of such models and accumulates the interestingness of each model into a consolidated model. In order to consolidate the models, the system summarizes the associations between the models in the collection and identifies support (models reinforce each other), complementary (models complement each other), and overlap of the models. The accumulated interestingness keeps track of historical exploration and helps the user summarize their findings which can lead to new discoveries. This mechanism for integrating results from multiple models can be applied to a wide range of decision support systems. We demonstrate our system in a case study involving the financial status of US companies.

**Keywords:** multiple models, model consolidation

## 1. INTRODUCTION

Visual data analysis is analogous to a treasure hunt. It is a process of identifying models that contain interesting information. A model, in our context, is an abstract representation of all or parts of the data. Each model provides clues or evidence as to certain patterns (e.g. anomalies and clusters) contained within the data. The final description of the contained information is usually the result of a progressive refining process referred as Exploratory Data Analysis (EDA). One big challenge that most systems/algorithms that support EDA[1–4] face is to reconstruct "a whole story" in the original data contents from multiple visual displays that are generated by some transformations. Linking the views,[5–7] obviously, is a ready solution to this challenge, however, most linking methods do not support extended connections among all the historical views generated during the exploration. John Tukey pointed out:[8] *" Ideas come from previous exploration more often than from lightning strokes"*. Motivated by his words, we believe that collecting and maintaining previous findings, undoubtedly, can lead to new discoveries.

Our goal is to develop a system that facilitates a progressive consolidation of the multiple data models generated in an explorative process. A single model often provides limited evidence of certain patterns among the data items, based either on a subset of the data dimensions, some dimensionality reduction or projection process designed to convey high dimensional relations, or computational models derived from all or parts of the data. Subsequent models can support this evidence, argue against the hypothesized relationship, or provide evidence for a totally new hypothesis. Our system builds a framework for combining these disparate models into a more complete and confident description, creating "a whole picture" of the original data contents.

We introduce our concepts via an example. Sally wants a new car and she may have certain requirements: A: (budget $<=$ \$ 20$k$), B: (6 cylinder engine), C: (hwy mpg $>=$ 30), D: (horse power $>=$ 400). Unfortunately, she cannot find any car that satisfies all her needs, and she has to explore the market a bit more by trying combinations of these conditions. The number of combinations grows exponentially as more and more conditions are considered. In our approach, we think of the 4 conditions as 4 models, and illustrate the model associations before further investigation; for instance, B and D are two models that overlap with each other and can be consolidated to a new model. Given the model associations and model consolidations, progressively adding models (e.g. F:(budget $<=$ \$ 30$k$)) becomes a more manageable explorative strategy compared to exhausting possible combinations.

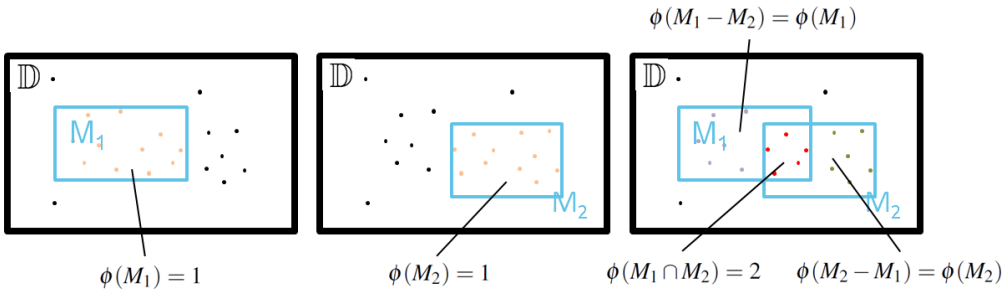


Figure 1: Example of model consolidation and the definition of the consolidation operator $\phi$

The major contributions of this work are: 1. We provide a framework that allows the user to summarize historical models that are generated during the exploration. The consolidated model can be either used as evidence of some hypothesis, or further combined with other new models; 2. We implement a model association view, which depicts the relationships between the models in the historical collection. This interface allows the user to interactively build consolidated models by choosing from the collection based on the model relations.

The remainder of this paper is organized as follows: In section 2, we discuss existing techniques that are related to ours. Section 3 formalizes the problem we are solving. The visual representation of the system is demonstrated in Section 4. Sections 5 presents an application of our system on a financial dataset. Finally, we sumarize our work and discuss future directions in Section 6.

## 2. RELATED WORK

Many terms have been used to describe data analysis and knowledge discovery that is done in stages, such as *incremental*, *progressive*, and *ensemble-based*. The notion of incremental data mining has primarily focused on adapting models when new data arrives,.[9,10] The key idea of these streaming-data-driven mining strategies is to automatically detect changes of the data and adjust the models dynamically along with the changes. Another type of incremental modeling is to maintain the extracted patterns[11] for historical pattern analysis.

Progressive mining, on the other hand, often refers to the notion of performing analysis at multiple granularities, which is usually a user-driven process. DBMiner,[12] for example, use the terms *progressive deepening* and *progressive generalization* to indicate drilling down or rolling up within data cube representations to change the resolution of analysis. Another progressive refinement system[13] uses multiple resolutions within multimedia data to extract content-based associations.

The concept of ensemble learning[14] is perhaps most relevant to our current work. This involves utlizing multiple models to analyze a dataset. EnsembleMatrix[15] is an interactive tool that allows the user to build customized classifiers out of multiple trained classifiers. Confusion matrices are used to identify the strengths and weaknesses of individual classifiers, and a barycentric coordinates display is used to specify weightings for all the classifiers in generating the final classification.

Other work has inspired our research as well. Yang et al.[16] defines a Nugget as a piece of interesting information embedded in the dataset. A management system was designed and implemented to refine and maintain the nuggets. The system *Prospect*[17] demonstrates a method that aggregates a collection of classification models. The aggregated results are used to debug the models.

## 3. PROBLEM DEFINITION

During an EDA process, models generated by the user may or may not contain information the user wanted the most. To achieve better discoveries the user usually has to adjust the strategies based on the current less interesting models. The adjustments, however, mostly rely on the users' memory of the explored models. The key questions are:

*What are the connections between the models that are generated so far?*

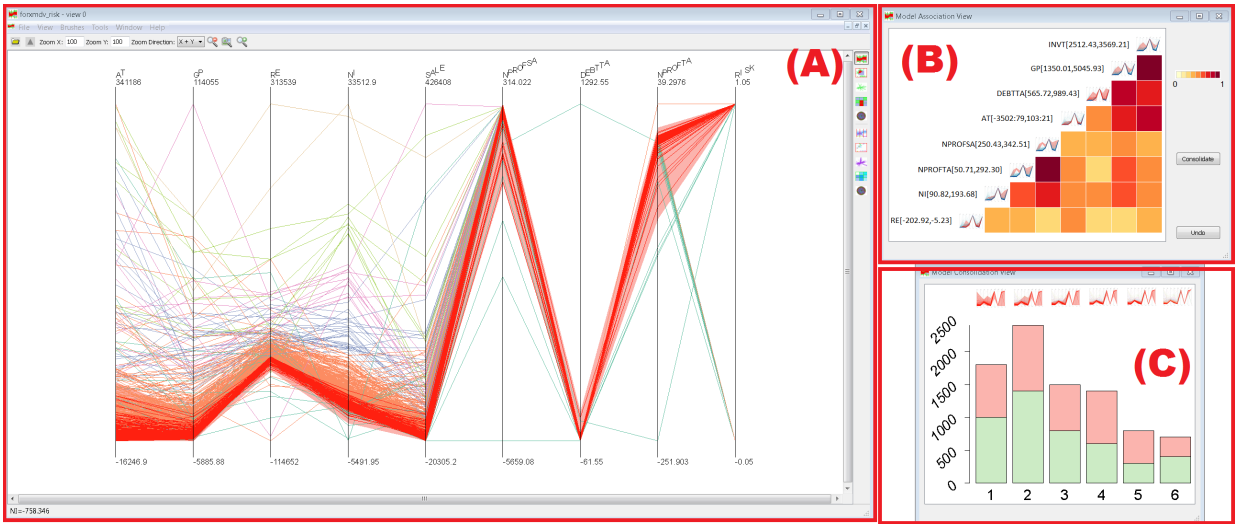


Figure 2: The interface for data model consolidation. (A) data view, where the user can define various models, the polylines are colored based on the values of the leftmost axis; (B) model relation view, where the user examines the relations between the models she creates; (C) consolidated models, where the user examines the coverage of models she consolidates.

*How do we utilize the explored models systematically to make discoveries?*

Our work attempts to answer these questions. We first formalize the problem description. Generally speaking, a model generated during an explorative process has a corresponding subset of data instances, and these can be conveyed in a visualization system, such as Parallel Coordinates, Scatterplots and Glyphs.[2] Since the visual representations can be easily constructed from the underlying data, we can define the model using the underlying data.

Let $\mathbb{D}$ be the dataset to be analyzed. Let $M$ be any model of $\mathbb{D}$ ($M \subset \mathbb{D}$. Note, intersection or union of models are still models). Let $\mathbb{M}$ be the collection of models and $\mathscr{R}$ be a symmetric relation on $\mathbb{M} \times \mathbb{M}$. Let $\mathbb{R} : \mathscr{R} \to R$ be a function that reflects the associations of the two models:

$$\mathbb{R}(M_1, M_2) = \frac{|M_1 \cap M_2|}{|M_1 \cup M_2|} \text{ where } M_1, M_2 \in \mathbb{M}$$

For example, let $M_1 = \{1,2,3\}$ and $M_2 = \{2,3,4\}$, $\mathbb{R}(M_1, M_2) = \frac{|\{2,3\}|}{|\{1,2,3,4\}|} = \frac{2}{4} = 0.5$

In order to summarize the previous findings and combine the historical models, we define an operator $\phi : M \to N$ (where $M$ is a model and $N$ stands for natural numbers) that reflects the interestingness of combined models (e.g. $M_1 \cap M_2$). We define this operator based on the assumption that each model has a certain degree of interest. Without specifying, we assume the default value of the interestingess is 1 for any simple model $M$ (model that is not generated by consolidation process, thus $\phi(M) = 1$). Since each model partitions the data space differently, the coverage of multiple models can create complex boundaries. In order to characterize the interestingness of data points covered by multiple models, we define the consolidation process as follows. Given two simple models $M_1$ and $M_2$, we have $\phi(M_1 \cap M_2) = \phi(M_1) + \phi(M_2)$; $\phi(M_1 - M_2) = \phi(M_1)$; $\phi(M_2 - M_1) = \phi(M_2)$ (shown in Figure 1). Basically, this operator accumulates the interestingness of data points that appears in multiple models during the explorative process. After we consolidate the two models $M_1$ and $M_2$, we replace all the relations regarding $M_1$ or $M_2$ with new relations regarding the new model $M_{1 \cap 2}$ so that we can continue the consolidation process progressively with updated model associations.

The model associations and the consolidated summary are two types of important information we want to deliver to the user and we discuss the design of our views regarding such information next.

## 4. SYSTEM DESCRIPTION

In this section we describe each of the components of our visual design for the two types of information we defined in the previous section. While each type of information can be represented in a number of ways, we have selected methods we believe are reasonable, and concentrate on linking all parts into a complete pipeline.

### 4.1 Model Association View

A Venn diagram is perhaps the most straightforward visual technique to present the associations among several sets.[18] However, the complexity of Venn diagrams grows dramatically when the number of sets is over 4 (due to the fact that the combinations of sets grows exponentially; a 7 set diagram can be seen at http://moebio.com/research/sevensets/). The problem we are trying to solve in this section is how to present multiple-model relationships in a straightforward way as in Venn diagrams, and scaling better than Venn diagrams.

One way of solving this problem is to only show the relation between each pair of models in a single view. This strategy avoids the multiple-model combination problem, but creates a new problem. How do we recover the information about the multiple model combination? The answer is that we allow the user to progressively consolidate models step by step; thus the multiple model relations can be converted to a sequence of two-model relations.

The design of the view is similar to the the *Collect:Pearson Operator* view in,[19] and also similar to the scatterlots score view in.[20] Instead of showing the relation of two variables or the scores of scatterplots, we show the relation of two models: $\mathbb{R}(M_1, M_2)$. In this view, we arrange the user generated models $\{M_1, M_2, \ldots, M_n\}$ on the diagonal of the display where thumbnails of the models are visualized. The user can create models via brushing operators in any visualization representations or via any computational techniques. Without loss of generality, we mainly use parallel coordinates in our examples. The thumbnail, which is a snapshot of an original model, links to the original model view at a higher resolution. The graph below the diagonal is used to visualize the relations between each pair of models. Each position $(i, j)$ corresponds to the value of a relation $\mathbb{R}(M_i, M_j)$. The graph above the diagonal is used for model annotations. The annotation of a model characterizes how the model is generated by the user, for instance, brushing over attribute $X$ with range $[a, b]$. Figure 3 and Figure 4 shows the design of the *Model Association View*. We show the scalability of our system in Figure 4 which displays the relations between 30 models. The next subsection describes our visual design for the *Model Consolidation View*.

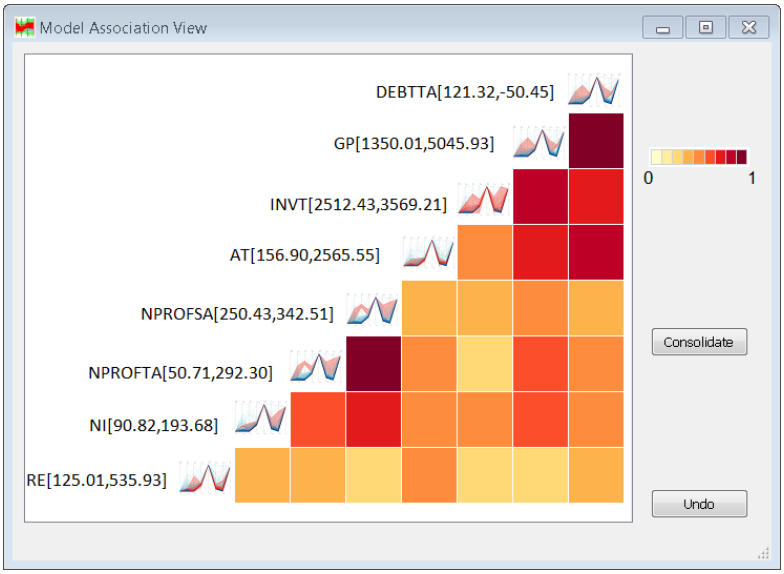


Figure 3: The *Model Association View* displays pairwise relations between 8 models. The annotation and the thumbnail of a model indicates how it is created. The thumbnails on the diagonal indicates the corresponding model of any color cell. A color cell indicates the relation value of two models using a color range from red ($\mathbb{R} = 1$, two models have much in common) to light yellow ($\mathbb{R} = 0$, two models have little in common).

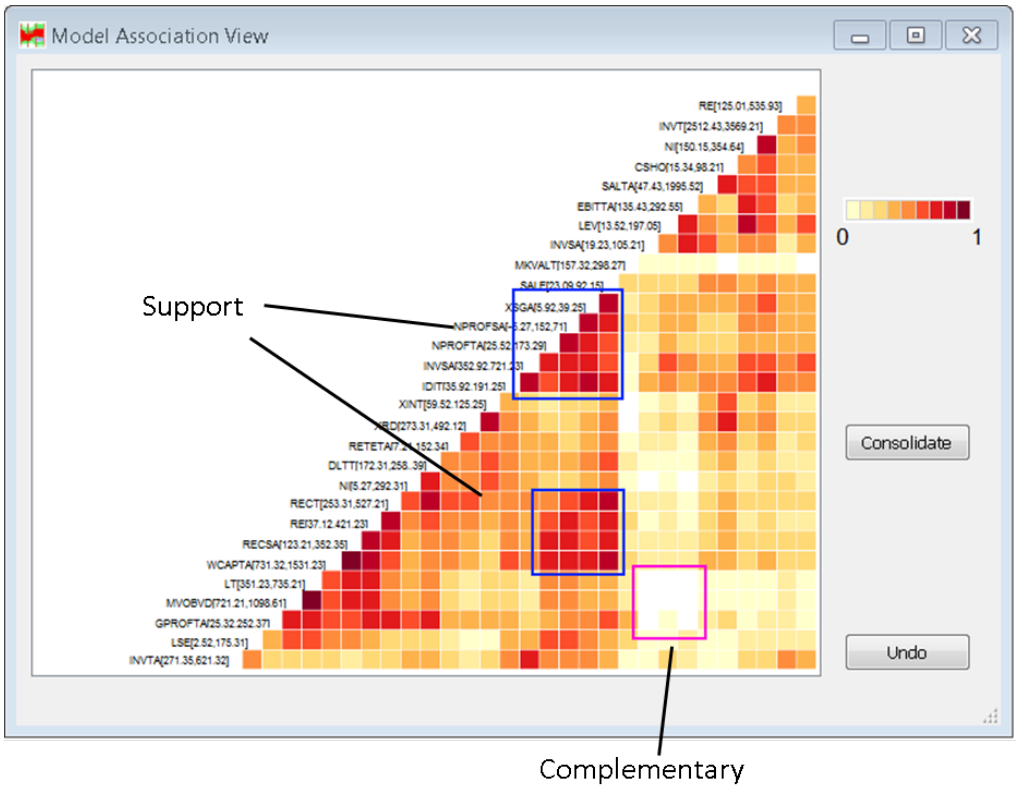


Figure 4: *Model Association View* shows 30 models browsed by the user during an explorative process. The user is able to identify regions of models that support each other and the regions of models that complement each other.

### 4.2 Model Consolidation View

As we described in Section 4.1, we trade the full combinations of multiple models for scalability. Therefore, we need to summarize the two-model consolidation at each step using the operator $\phi$ defined in Section 3. The purpose of this view is to visualize the summary of the consolidation operations by the user. The information we want to deliver is the value $\phi$ and the corresponding models that contribute to that value. However, continuously consolidating models produces a very large number of possible sub-models (subsets of models), most of which may be very small or even empty. To make it more scalable, we merge the sub-models based on $\phi$. After the merging, the number of sub-models is bounded to the value of $\phi$ that is determined by the number of consolidations the user performs. With the summary information, the user can tell how the consolidation works, what portions of the models contribute the the new model, and what are the proportions of the consolidated model at each level of $\phi$. The proportion of models contributing to the new model tells the user, which model plays a more important role during the consolidation.

Now we describe how we designed the *Model Consolidation View* (Figure 5) based on the above considerations. We use a stacked barplot to illustrate the merged models at each level of $\phi$. Each vertical bar in the plot stands for part of the consolidated model that is summarized at some level of $\phi$, where the height indicates the size (number of data points covered by the model) of the model. We also keep track of the last two models used to form the current consolidated model. The two models are characterized by the upper and lower bars of the stacked barplot, respectively. The ratio between the two stacked bars is determined by the contribution of the two models to the consolidated model. For example: the consolidation of $M_1 = \{1,2,3\}$ and $M_2 = \{2,3,4,5\}$ results in a new model of $\{\{2,3,\},\{1,4,5\}\}$ where $\phi(\{2,3\}) = 2$ and $\phi(\{1,4,5\}) = 1$. The contribution ratio for $M_1$ and $M_2$ regarding sub-model $\{2,3\}$ is

$$4/3 = \frac{|\{2,3\}|/|\{1,2,3\}|}{|\{2,3\}|/|\{2,3,4,5\}|}$$

The sequence of consolidating models does not affect the resulting model, as long as the chosen models are the same. This is anlogous to the Venn diagrams that the order of combining sets does not affect the final visualization. To facilitate the

understanding of the consolidated model, thumbnails of the sub-models at each level of $\phi$ are also included in this view. The user can go to the original scale by clicking on any of them.

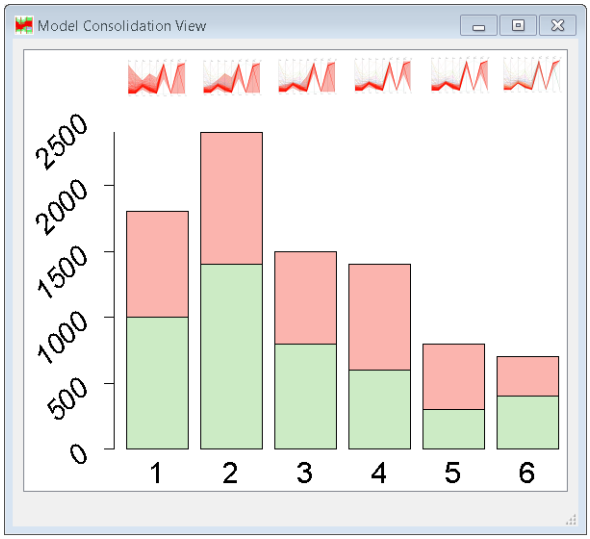


Figure 5: The *Model Consolidation View* displays a progressively consolidated model by the user over 6 steps. The y axis indicates the size of the model (number of data instances covered by the model) at each level of $\phi$. The x axis is the accumulated value of $\phi$ over the course of progressive consolidation. The top line of thumbnails are visual representations of the consolidated models. The top portion and the bottom portion of the stacked barplot characterize at what ratio the two previous models contribute to the current consolidated model, which is represented by the total height.

## 5. APPLICATION

Now we show the analytical capabilities of our proposed work using a case study based on a financial dataset. The dataset (44 dimensions and 4013 data instances) is extracted from Compustat (http://www.compustat.com/); which is commercial database. A domain expert (Professor of Finance from School of Business) participated in our case study and supervised the process. The task of our study was to refine the criterion that indicates the companies at risk of bankruptcy. We started with 8 typical financial variables as the expert suggested, and created models based on domain knowledge; for example, one model can be the companies with low asset values denoted as $AT : [-3502.79, 103.21]$.

First, we established an experiment with a traditional parallel coordinates view. The expert started to try the combinations after examining each individual model. Through the think-aloud session, we learned that the combinations she made are primarily based on her experience and the shape of parallel coordinates of the individual models. She admitted that the characteristics of a sample of dataset can vary a lot due to many factors, for example, type of industries, development levels of the economy and regional policies. Therefore, using experience to combine the individual models can be tricky.

Then, we showed our system to the expert. After she understood the representations of the three views, she thought the Model Association View was helpful for her to refine the models she created. She also got a general idea of how to combine the models to achieve a consolidated model with proper coverage of the dataset.

Now we describe how the system works, step by step. The first step is to identify interesting models in a multivariate analysis display. The user browses the data space at this step, and marks the models of interest in the data view (Figure 2 (A)). All the marked models are visualized in the model relation view (Figure 2 (B)) where the user learns the relations between the models she generates. The second step is to combine models progressively based on the model relations. In Figure 3 we notice that the model GP[1350.01,5045.93] (gross profit) and the model INVT[2512.43,3569.21] (Inventory) share a large portion of data, as indicated by the dark red cell. The expert proceeded with the selection of this cell and combined the two models. By combining the models with such a pattern (dark red color), the intersection of the two models still has sizeable coverage compared to the original models and satisfies both ranges. The third step is to examine the consolidated model in Figure 5. Sometimes, the coverage of the consolidation model with highest $\phi$ value is small

(model with $\phi = 6$). If the user prefers a model with a larger coverage, she can stick to the model with a smaller $\phi$ value, which has a relaxed combination of the input models and often larger coverage.

The consolidation result of 8 models is shown in Figure 5. From the highest $\phi$ value ($= 6$), we can tell, there exist companies in the data sample that can be covered by a maximum of 6 of the compound models and no companies can be covered by more than 6 models. The size of the subset of the companies covered by 6 models is about 800. The value $\phi$ is restricted by the number of bins that can be displayed in a histogram. The bins listed various alternative models that are generated during the consolidation process. Additionally, the data view of the subset can be seen in Figure 2 (A), where the band highlighted in red shows the model coverage. The highlighted portion of data is covered by multiple models where each model conresponds to one model previously created by the user. The resulting model covers the portion of data of most interest to the user. In our application, it is the portion of companies revealing high risks that are characterized by multiple constraints created by the user.

## 6. SUMMARY AND FUTURE WORK

We demonstrated a model consolidation system that makes use of all interesting models the user creates during an exploratory data analysis process. We designed and implemented a *Model Association View* that illustrates the model relations between user defined models, including progressively consolidated models. We also developed a *Model Consolidation View* that visualizes the new models the user creates. Through an expert interview session, we showed the effectiveness of our system compared to the classical multivariate visualization methods (e.g. parallel coordinates).

The work we presented in this paper has several limitations. Some of the limitations lead to interesting new directions. In this section, we discuss two open problems we encountered during the design and implementation of our system. We also discuss the weaknesses of our system that were pointed out by the expert during our case study, including the design of the interface and missing features.

**Model Summarization Granularity:** Our current solution offers a coarse resolution of the consolidated model space. There are still meaningful questions we cannot answer with our current solution. For example, when the user identifies a consolidated model covered by $n$ out of $N$ input models, she asks which $n$-model combinations are equivalent to this consolidated model. Answering this question can help the user further refine the input models and make the consolidated model more interpretable. Take our case study as an example; $\phi = 6$ means the model space is covered by 6 other models. The coverage consists of $\binom{8}{6} = 48$ combinations. This problem is not quite scalable, because in general, when a model space is covered by $n$ models, the possible number of ways is $\binom{N}{n}$. Converting this problem to a visual pattern recognition problem seems a plausible solution to us and we are still investigating it.

**Model Substitution Problem:** While watching the expert create models during the exploration, we figured model substitution could be useful. Along with the consolidation process, more and more complex consolidated models are created from simpler ones. Although they are representing the proper subset of data instances the user is interested in, the annotation of such complex models is not an easy task. Substituting a complex model with a relatively similar and simpler model becomes an intuitive solution to more and more complex models. Take $\phi = 6$ as an example again. If we are able to substitute it with a much smaller number (say, 2) of other models, the interpretation of such a model will be greatly improved. Currently, our technique provides the graphical representation of such models with no annotations; we are considering adding substitution and interpretable annotations in the near future.

**Usability Improvement:** We also collected feedback from the expert during an interview. She made several suggestions about our system regarding the user interface and functionalities. First, she thought more information about the models should be presented, such as mean and variance. She requested this because most analysts in her domain usually use such statistics to describe a model. Second, she felt Venn diagrams are more intuitive in characterizing model consolidation. Although it is not scalable, showing how the two models are merged at each step is still feasible. Third, she thought a preview of the two models in the data space can additionally help the consolidation step. Compared to the other two directions, this is more about enhancing the component we already completed than a matter of how to visualize model consolidations intuitively. As we already demonstrated, we have a pairwise solution to this problem. Is there any other way of showing the multiple model relationships? It may be another interesting but very challenging question.

## 7. ACKNOWLEDGEMENT

This work is supported under NSF grant IIS 1117139.

## REFERENCES


[1] Swayne, D. F., Lang, D. T., Buja, A., and Cook, D., "Ggobi: Evolving from xgobi into an extensible framework for interactive data visualization," *Computational Statistics & Data Analysis* **43**(4), 423–444 (2003).

[2] Ward, M. O., "Xmdvtool: Integrating multiple methods for visualizing multivariate data," in [*Proceedings of the Conference on Visualization'94*], 326–333, IEEE Computer Society Press (1994).

[3] R Development Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2008). ISBN 3-900051-07-0.

[4] Friedman, J. H. and Tukey, J. W., "A projection pursuit algorithm for exploratory data analysis," *IEEE Transactions on Computers* **100**(9), 881–890 (1974).

[5] Piringer, H., Kosara, R., and Hauser, H., "Interactive focus+ context visualization with linked 2d/3d scatterplots," in [*Coordinated and Multiple Views in Exploratory Visualization, 2004. Proceedings. Second International Conference on*], 49–60, IEEE (2004).

[6] Roberts, J. C., "Exploratory visualization with multiple linked views," in [*Exploring Geovisualization*], MacEachren, A., Kraak, M.-J., and Dykes, J., eds., Amsterdam: Elseviers (December 2004).

[7] Jern, M., Johansson, S., Johansson, J., and Franzen, J., "The gav toolkit for multiple linked views," in [*Coordinated and Multiple Views in Exploratory Visualization, 2007. CMV '07. Fifth International Conference on*], 85–97 (2007).

[8] Tukey, J. W., "We need both exploratory and confirmatory," *The American Statistician* **34**(1), 23–25 (1980).

[9] Kifer, D., Ben-David, S., and Gehrke, J., "Detecting change in data streams," in [*Proceedings of the Thirtieth international conference on Very large data bases - Volume 30*], *VLDB '04*, 180–191, VLDB Endowment (2004).

[10] Yang, D., Guo, Z., Rundensteiner, E. A., and Ward, M. O., "Clues: A unified framework supporting interactive exploration of density-based clusters in streams," *20th ACM Conference on Information and Knowledge Management* , 815–824 (2011).

[11] Yang, D., Rundensteiner, E. A., and Ward, M. O., "Summarization and matching of density-based clusters in streaming environments," *Proceedings of the VLDB Endowment* **5**(2), 121–132 (2011).

[12] Han, J., Fu, Y., Wang, W., Chiang, J., Gong, W., Koperski, K., Li, D., Lu, Y., Rajan, A., Stefanovic, N., Xia, B., and Zaiane, O. R., "Dbminer: A system for mining knowledge in large relational databases," in [*Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*], 250–255, AAAI Press (1996).

[13] Zaiane, O. R., Han, J., and Zhu, H., "Mining recurrent items in multimedia with progressive resolution refinement," in [*Data Engineering, 2000. Proceedings. 16th International Conference on*], 461–470, IEEE (2000).

[14] Zhou, Z.-H., [*Ensemble methods: foundations and algorithms*], Chapman and Hall/Crc (2012).

[15] Talbot, J., Lee, B., Kapoor, A., and Tan, D. S., "Ensemblematrix: interactive visualization to support machine learning with multiple classifiers," in [*Proceedings of the 27th international conference on Human factors in computing systems*], *CHI '09*, 1283–1292, ACM, New York, NY, USA (2009).

[16] Yang, D., Rundensteiner, E. A., and Ward, M. O., "Analysis guided visual exploration of multivariate data," in [*Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*], 83–90, IEEE (2007).

[17] Patel, K., Drucker, S. M., Fogarty, J., Kapoor, A., and Tan, D. S., "Using multiple models to understand data," in [*Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Two*], 1723–1728, AAAI Press (2011).

[18] Wilkinson, L., "Exact and approximate area-proportional circular venn and euler diagrams," *Visualization and Computer Graphics, IEEE Transactions on* **18**(2), 321–331 (2012).

[19] Ingram, S., Munzner, T., Irvine, V., Tory, M., Bergner, S., and Möller, T., "Dimstiller: Workflows for dimensional analysis and reduction," in [*Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*], 3 –10 (2010).

[20] Seo, J. and Shneiderman, B., "A rank-by-feature framework for interactive exploration of multidimensional data," *Information Visualization* **4**(2), 96–113 (2005).