Diverse Information Integration and Visualization

Susan L. Havre, Anuj Shah, Christian Posse, Bobbie-Jo Webb-Robertson
Pacific Northwest National Laboratory, P.O. Box 999, Richland, WA 99352

ABSTRACT

This paper presents and explores a technique for visually integrating and exploring diverse information. Researchers and analysts seeking knowledge and understanding of complex systems have increasing access to related, but diverse, data. These data provide an opportunity to consider entities of interest from multiple informational perspectives not available from any single, data or information type. These multiple perspectives are derived from diverse, but related data and integrated for simultaneous analysis. Our approach visualizes multiple entities across multiple perspectives where each perspective, or dimension, is an alternate partitioning of the entities. The partitioning may be based on inherent or assigned attributes such as meta-data or prior knowledge captured in annotations. The partitioning may also be directly derived from entity data; for example, clustering, or unsupervised classification, can be applied to multi-dimensional vector entity data to partition the entities into groups, or clusters. The same entities may be clustered on data from different experiment types or processing approaches. This reduction of diverse data/information on an entity to a series of partitions, or discrete (and unit-less) categories, allows the user to view the entities across diverse data without concern for data types and units. Parallel coordinate plots typically visualize continuous data across multiple dimensions. We adapt parallel coordinate plots for discrete values such as partition names to allow the comparison of entity patterns across multiple dimension for identifying trends and outlier entities. We illustrate this approach through a prototype, Juxter (short for Juxtaposer).

Keywords: **information visualization, visual analysis, categorical data, parallel coordinates**

1. INTRODUCTION

Many of the grand unsolved problems are very complex and multifaceted. It's hard to imagine that any single experiment, data set, or analysis tool could consistently ferret out a terrorist cell or tell us how to cure cancer. For complex problems we need to leverage a variety of resources. In systems biology, there are new, exciting high-throughput technologies that measure various features of experimental samples. The raw, numeric data produced are often analyzed by methods developed for a specific type of data. While the opportunity exists to measure several different properties of the same sample, the challenge remains to integrate the diverse data results for analysis that spans the various properties simultaneously. To further complicate matters, biologists want to analyze their experimental data in the context of curated information accumulated in public (and private) databases.

Many fields are experiencing the same or similar challenges. Our immediate research is focused on integrating diverse biological data though we believe our technique is broadly applicable. We observe that much of the annotation data is categorical, for example, metabolic pathways, biological functions, or cellular location of proteins. Numeric data is often categorized as well; for instance, a user might partition, or bin, genes into three gene expression groups, up, down, or no change based on a gene expression ratio. Clustering, or unsupervised classification, is a more sophisticated approach to aggregation by similarity that utilizes the multi-dimensional numerical data of entities to partition them into clusters of similar entities. Cluster membership is a derived categorical attribute of an entity. The same entities may be clustered on data from different experiment types or processing approaches. The ability to represent diverse data as categorical information allows the integration of a potentially very mixed bag of data. The data integration is achieved by depicting the entities across all the categorical dimensions in the same graphic. Users can look for patterns suggesting relationships across information derived from experimental results, meta-data, and annotations.

Each partitioning of a data set defines a dimension with discrete values; the discrete values are text strings that may be meaningful names, imply an order, or simply distinguish partitions. For example, documents in a document collection might each be characterized by cluster membership, user-defined groupings (based on query results), date published, document source (e.g., news feed or email), and source location; the documents could be viewed across all five

Visualization and Data Analysis 2006, edited by Robert F. Erbacher, Jonathan C. Roberts, Matti T. Gröhn, Katy Börner,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 6060, 60600M, © 2006 SPIE-IS&T · 0277-786X/06/$15

dimensions. In systems biology, genes can be categorized with clustered gene expression data, associated protein quantities, metabolic pathways, and cellular locations. Derived data may have associated metrics such as confidence scores that can also be binned and viewed as a separate dimension parallel to the derived data dimension. Parallel coordinates plots can effectively visualize continuous data across multiple axes, or dimensions. We prefer an approach that will allow viewing the data with respect to multiple dimensions rather than a series of pair-wise comparisons or a dimension-reduced view. However, there are obstacles to visualizing categorical data effectively in a parallel coordinates plot.

In this paper we present an approach for adapting parallel coordinate plots to explore data across multiple categorical dimensions. We discuss related work including the static precursor of our approach, parallel coordinate plots, and visualizations designed specifically for categorical data. We then describe the contribution of our work. Our approach is illustrated through our prototype, Juxter, followed by a discussion of things we have learned about this approach while implementing and applying it to biological data. We conclude with a summary and intended future work.

2. RELATED WORK

Our work is related to a diverse collection of previous work including a static illustration of gene expression data, parallel coordinates, and the visualization of categorical data.

2.1. Inspiration

Our work is based on an illustration that appeared in two papers by Carr et al. [1] and Michaels et al. [2]. The illustration shows the clustering of gene expression data based on two similarity measures: mutual information and Euclidean distance, described in the papers. The data from the papers are depicted in a similar layout by Juxter in Figure 1. The two clusterings (partitionings) are represented by horizontal axes labeled Euc_Distance and Mutual_Info to the left of the figure. The clusters of each clustering are positioned along their respective axis. A poly-line representing each gene is drawn vertically through its assigned cluster along each axis. The lines extend to include initial and final points along additional axes (top and bottom) indicating gene family and protein function, respectively. A viewer can track correlated line patterns to discover trends and outliers.

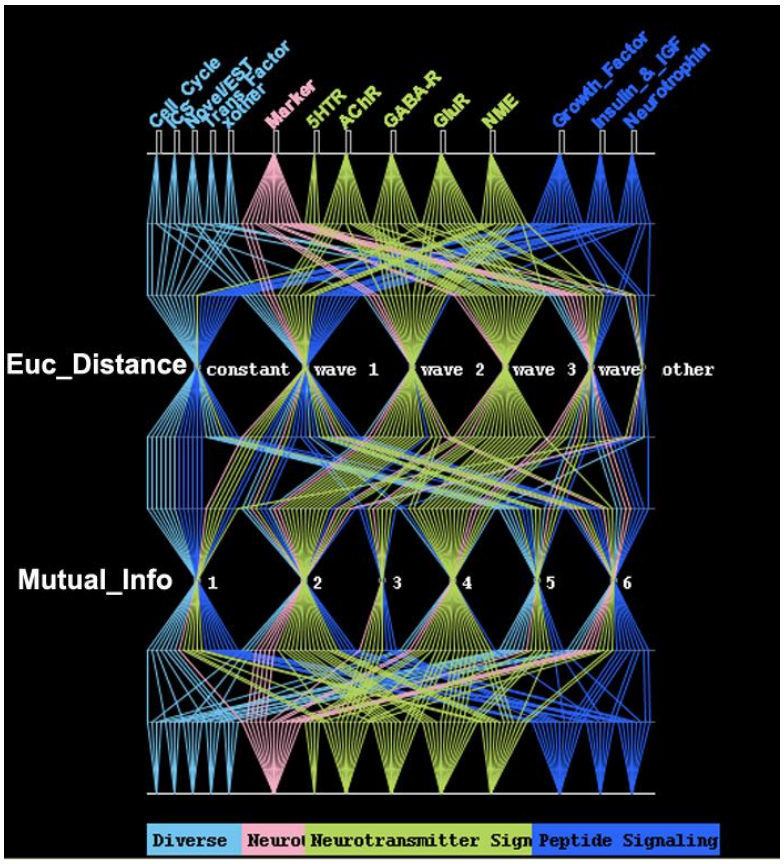

Figure 1: Gene expression data clustered by two different similarity measures. Each vertical line shows the path of a gene across these clusterings. The genes are grouped at the top and bottom by gene family and associated protein function, respectively.

Observe the flow of yellow lines representing genes in the *Neurotransmitter Signaling* functional group (at the bottom). These all map back to the group of gene families with yellow labels at the top. By tracking the flow of yellow lines, it is clear that most of these genes belong to *wave 2* and *wave 3* on the Euc_Distance axis and clusters *2* and *4* on the Mutual_Info axis. The cyan genes in

the *Diverse* functional group belong mostly to the *constant* cluster on the Euc_Distance axis. These genes are split between clusters *1* and *5* on the Mutual_Info axis.

Clearly, the two clusterings are not identical, but the colored lines indicate some correlation between them. Observe that there is also some correlation between the clusterings and the categorical groupings at the top and bottom. These annotations, gene family and functional role at the top and bottom, respectively, provide valuable contextual information that helps characterize the clusters. Carr et al [1] provide an interesting discussion on cluster comparison and some of the design decisions behind the illustration. The generalization and extension of this work as a dynamic, interactive visualization was not pursued by Carr et al largely due to the complexity in calculating an optimal layout. To our knowledge, layout details for the original illustration such as gene, cluster, and clustering order were determined specifically for the data set and external to the graphic software.

2.2. Parallel Coordinates and Discrete Data

The illustration in [1, 2] is a variation on parallel coordinate plots first published by Inselberg [3, 4] and extended by Wegman [5] and others. Parallel coordinate plots can depict large amounts of high-dimensional, continuous (numeric) data. Each dimension is represented along an axis; all the axes are parallel to each other. There is no orthogonal axis and so no inherent or implied ordering of the parallel axes. The data along any parallel axis are typically continuous. For each element, a poly-line is drawn through the appropriate position on each parallel axis. Parallel coordinate plots of continuous data scale well to handle large numbers of elements, or entities, up to the point where occlusion from overlapping lines, or overdrawing, becomes a problem. The number of visible axes (dimensions) is limited by the display space.

Parallel coordinate plots do not scale as well with categorical data. Consider an axis line of 100 pixels. Theoretically, a continuous data point might fall on any one of the 100 pixels. However, a categorical data point can only fall on the pixel chosen to represent the category. If there are two categories, then only two axis pixels will be crossed. We assume that the number of categories is much smaller than the number of available pixels; otherwise, our data are effectively continuous. The reduction of points where lines may cross an axis reduces the available drawing space, and lines are much more likely to be overdrawn.

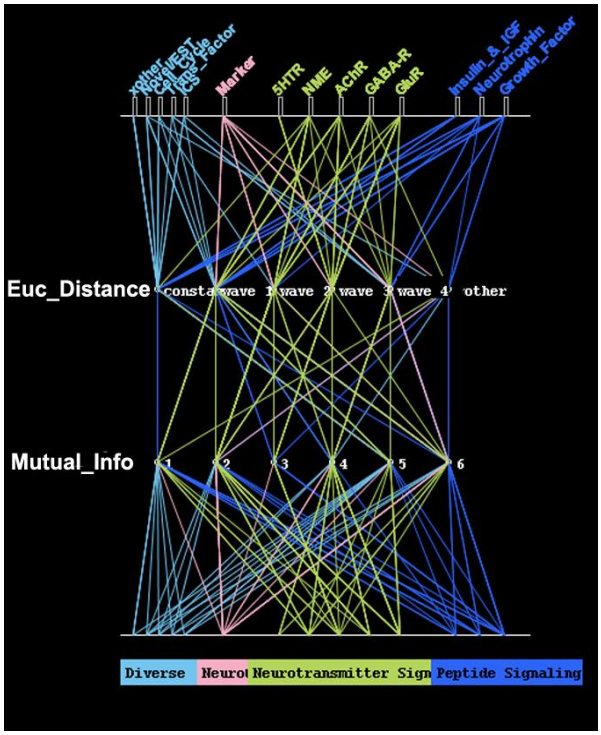

Figure 2: Juxter configured as a standard parallel coordinates plot showing the same data as Figure 1.

A more difficult problem with parallel coordinate plots of categorical data is the increased likelihood of multiple lines passing successively through the same points. If multiple items have identical values along neighboring axes, their lines will overlap exactly. This means that only the last line segment drawn will be visible. When such overdrawing occurs, valuable information is lost about which path an individual element takes across the axes and which other (and how many) items take the same path. In Figure 2, Juxter is configured as a parallel coordinate plot of the data in Figure 1; the difference between the two figures is in the line drawing between axes. The overdrawing of lines is clearly a problem in Figure 2. Individual genes are untraceable; for instance, none of the cyan genes can be traced from the Euc_Distance to the Mutual_Info clustering The makeup and relative size of the groups as well as the relationships among multiple genes are lost; for example, compare the information content in Figures 1 and 2 for genes that belong to both the *constant* cluster of Euc_Distance and the *1* cluster of Mutual_Info. Figure 1 reveals that these genes are split about equally between the *Diverse* and *Peptide Signaling* function groups. This cannot be seen in Figure 2.

However, overdrawing is not the only problem for parallel coordinate plots of categorical data. The placement of the points for the categorical values is also an issue. Because most of the categorical data points along each axis have no

inherent order, the placement of the points is necessarily arbitrary. Options for ordering dimensions and categories include: consult domain experts who might recognize an inherent order; apply optimal layout rules to minimize line crossings; or follow the order of dimension and group occurrence in the input file.

The work of Rosario et al. [6, 7] aims to resolve the problem of visualizing categorical data. They developed a technique, Distance-Quantification-Classing (DQC), for preprocessing nominal (categorical) data so that it can be viewed in a visualization designed for numeric data such as parallel coordinates. This preprocessing determines an order and relative spacing for the nominal values along parallel axes.

Graham et al. [8] introduced the notion of replacing the poly-lines with continuous curves in parallel coordinate plots that nicely handle dimensions of both continuous and discrete data. While the continuous curves follow the Gestalt principle of continuity much better than our jointed poly-lines, our approach better meets the proximity and similarity principles that assist the perception of trends.

2.3. Visualization of Categorical Data

Friendly [9-11] has a large body of work on the visualization of categorical data. "Mosaic displays" is perhaps the most well-known of Friendly's visualizations; it is based on the graphical metaphor of nested regions built on the notion of a data hierarchy. Friendly's data typically consist of counts of category populations rather than multi-faceted data about individuals (or entities) across multiple categorizations.

CatTree [12] visualizes categorical data in an extension of Treemaps. Treemaps [13] is a space-filling technique for visualizing hierarchical data. CatTree creates a hierarchy of the categorical data that can be visualized, queried, and re-ordered dynamically.

These and similar approaches assume a hierarchical ordering of data; the effectiveness of the visualization is heavily dependent on the ordering of the dimensions in the hierarchy. We don't want to be limited to hierarchical data or be required to force the data into hierarchies. At least initially, our approach avoids any such assumptions about the dimensions or partitionings.

3. CONTRIBUTION

We recognized that the complex, diverse data reformulated as categorical information can be integrated for directly observing patterns across the multiple data types. Our aim is to integrate very diverse biological data from multiple experimental techniques along with prior knowledge. Much of the state of the art high-throughput biological data requires preprocessing and analysis developed specifically for each experimental approach. The derived result is often identification, classification, and/or quantification of entities measured under careful experimental conditions. Most of this information, including the meta-data such as the experimental conditions, is naturally categorical; the rest can be binned or clustered to derive categorical information. By depicting an entity as a poly-line connecting the categories of that entity, we create a visual trace that supports the comparison of multiple entities over potentially very diverse data.

Further, we recognized that the line drawing technique in the illustration comparing two clusterings of gene expression data would generalize nicely to observing and comparing the patterns of entity traces. Patterns of correlation or outliers may be associated with biological meaning. We extend parallel coordinate plots, which support viewing patterns across multiple continuous dimensions, to handle categorical dimensions through a line-drawing algorithm that minimizes overdrawing. The line-drawing method aligns, at least in part, with a number of Gestalt principles to facilitate pattern recognition[14].

We developed an interactive prototype, Juxter, which dynamically creates the visualization from tabular data. We found that flexible, dynamic interactions reduce the need for complex optimization algorithms for graph layout. Using Juxter, researchers have visualized metagenomic information derived from contaminated soil; the researchers found biological insights they believe might have been missed, or taken longer to find, without Juxter [15].

4. VISUALIZATION TECHNIQUE

Although we developed our technique for integrating biological data, in this section we illustrate the visualization with a somewhat generic data set to avoid unnecessary domain-specific explanations. We selected the file BIOMED3.DAT from data sets offered online by the National Institute of Standards and Technology (NIST) at

http://www.itl.nist.gov/div898/software/dataplot/datasets.htm. This data set is a simple collection of categorical data. Although undocumented, it is sufficiently self-explanatory for our purposes. There are 193 records (patients) with several dimensions including *HOSPITAL, DATE*, and *AGE*; each of these dimensions has four values labeled *1, 2, 3*, and *4*, roughly four equal partitions. There are four unexplained dimensions named *X1, X2, X3,* and *X4*; each has values 1-4; we ignore these dimensions in this paper. A final dimension is labeled *CARRIER (0=NO)* where the group labels are 0 or 1. We replace the 0 label with *NO* and the 1 label with *YES*. While all the category labels are integers from 1 – 4, these are likely ordered nominal rather than numeric values. We assume that these data were collected to study the course of an outbreak and to characterize the carriers across hospitals, age groups, and dates as well as the mystery X1-4 dimensions.

4.1. Visualization

Figure 3 provides an overview of the data showing the patients as poly-lines across the Hospital (top), Date, Age, and Carrier axes. The Hospital axis is repeated at the bottom. The lines are colored to indicate the patient's hospital. One's attention is drawn to the larger areas of near-solid color. The cyan mass at the bottom left indicates that *Hospital 1* patients are mostly non-carriers and make up about one half of the *Age 1* group. We base this on the observation that almost all of the lines spreading up from *Hospital 1* link to the non-carrier group (see white arrow) and about half of the lines spreading down from the *Age 1* group are cyan (*Hospital 1*, see orange arrow). Also note that only one patient from *Hospital 1* is a carrier. This outlier can be seen as the single cyan line from the carrier group (*YES*) near the bottom right to *Hospital 1* at the bottom left (see red arrow).

Figure 4 shows the patients in *Hospital 1* selected. Selection of a line or groups of lines results in the non-selected lines being re-colored gray to effectively push them into the background. A user selects *Hospital 1* by clicking on the cyan label at the top or bottom, or rubber banding the cyan lines. The patients from *Hospital 1* are spread somewhat evenly across the four date groups. They clearly dominate in the first age group, but only a few of these patients belong in the *Age 4* group. Most notable, only one patient from *Hospital 1* is a carrier. This was noted in the discussion on Figure 3; however, it is much more obvious in Figure 4 (see red arrow).

In Figure 5, we select the *Date 1* group. This view

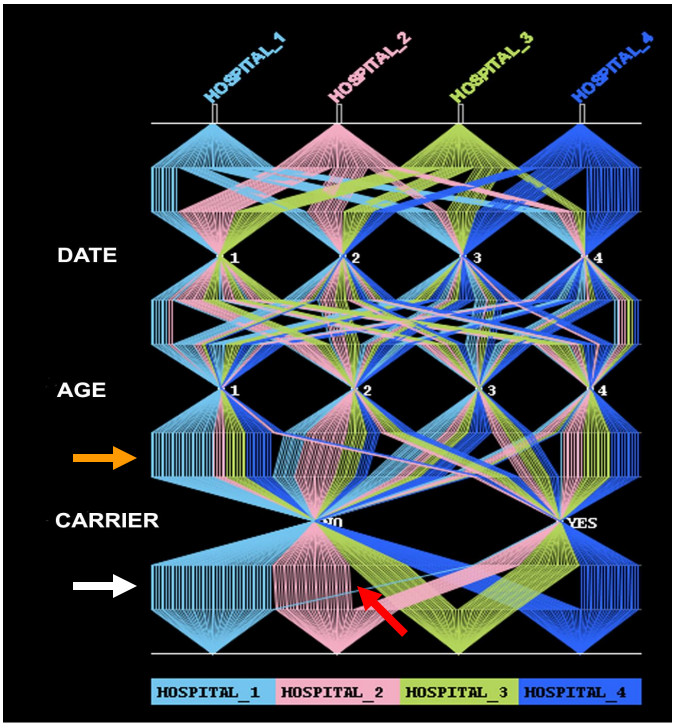

Figure 3: BIOMED3.DAT patient data across the following dimensions: Hospital (top), Date, Age, Carrier, and Hospital (repeated at the bottom.)

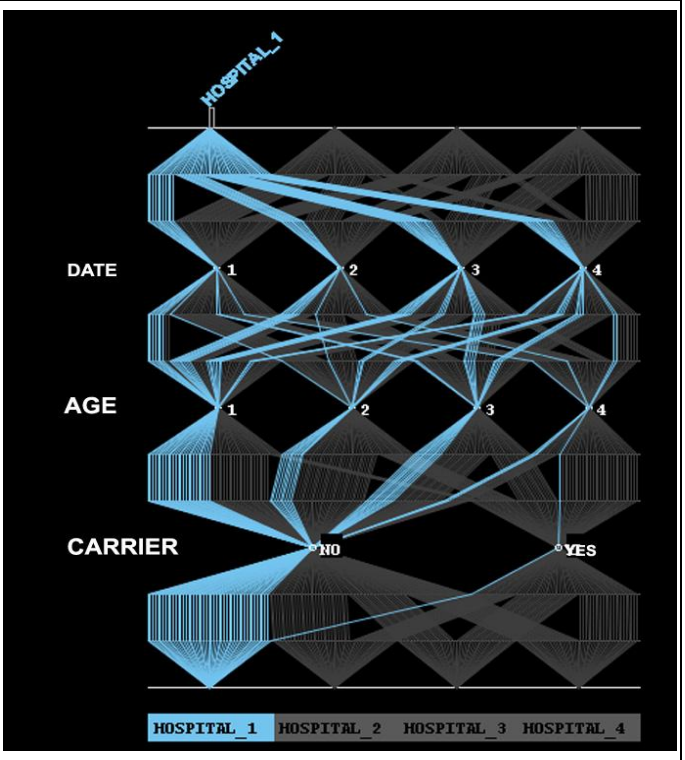

Figure 4: BIOMED3.DAT, same as Figure 3, but with the *Hospital 1* group selected.

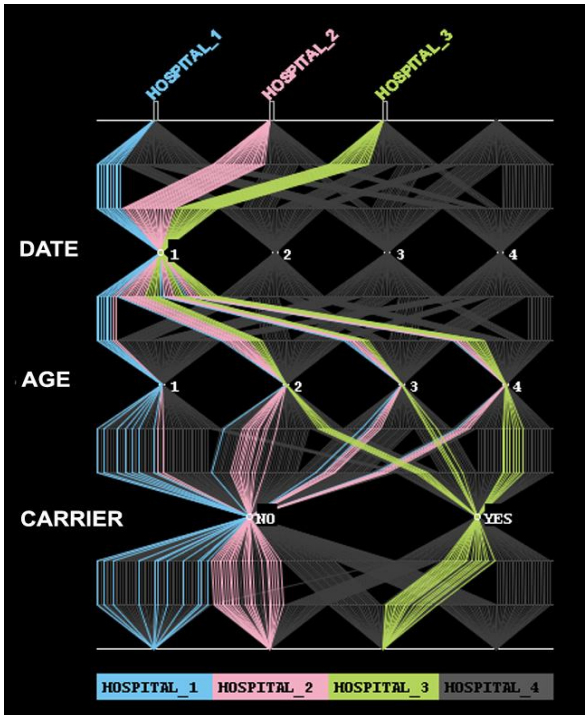

Figure 5: BIOMED3.DAT, same as Figure 3, but with the *Date 1* group selected.

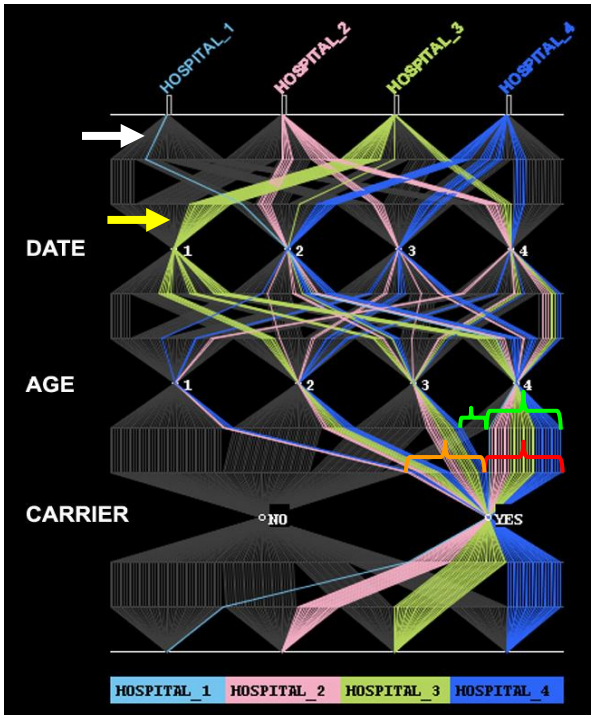

Figure 6: BIOMED3.DAT, same as Figure 3, but with the Carrier (*yes*) group selected.

clearly shows that none of the patients in the *Date 1* group are in *Hospital 4* (no dark blue lines or *Hospital 4* label). None of the patients from *Hospital 3* (green) in the *Date 1* group belong to the *Age 1* group. Interestingly, the *Date 1* group patients from *Hospital 3* are all carriers, while the ones from *Hospitals 1* and *2* are all non-carriers.

Figure 6 highlights the patients that are carriers. Again we see only one of these is in *Hospital 1* (white arrow) and all the *Date 1* carriers are in *Hospital 3* (yellow arrow). What we see clearly for the first time is that about half of the carriers are patients from the *Age 4* group (compare red and orange braces). Most of the *Age 4* group are carriers (green braces). The carriers are spread about equally across the four Date groups.

The observation of such patterns suggest hypotheses that upon further investigation may provide insights about relationships, for example, between hospitals and patient age, between dates of infection and hospitals. Possible hypotheses might be that (assuming that date and age groups are labeled to indicate increasing time and age) (1) *Hospital 1*'s patients were overall younger, non-carriers who fell sick in approximately equal numbers across the outbreak, (2) *Hospital 4*'s patents fell ill only during the last three periods and made up more than 1/3 of carrier group, (3) the carrier group tended to be older patients and, with one exception, were treated at *Hospitals 2, 3*, and *4.*

4.2. Line Spreading to Minimize Overdrawing

We have already mentioned that overdrawing is a big problem for parallel coordinate plots of categorical data. Line plots of categorical data have less available drawing area because the lines are constrained to cross axes only at limited, discrete points. In addition, data items may have matching group-to-group transitions across neighboring axes leading to overdrawing of lines. This reduced drawing area and increased risk of overdrawing lead to increased occlusion; the fewer categories per axis, the greater the occlusion problem. Occlusion means loss of information, thereby decreasing our ability to characterize individual items, categories, and the data as a whole. This is a lost opportunity to observe patterns, such as correlations or outliers in the collective paths across the categorizations.

To counter this problem, we spread the element lines by providing intermediate lines between the parallel axes. The user may choose between none, one, or two intermediate lines between each pair of axes. The data shown in Figure 3 (with two intermediate lines) are shown in Figure 7 with no intermediate lines. This view shows that at least one patient from each group on one axis transitioned to each of the possible groups in the next axis. The no-intermediate line view of data

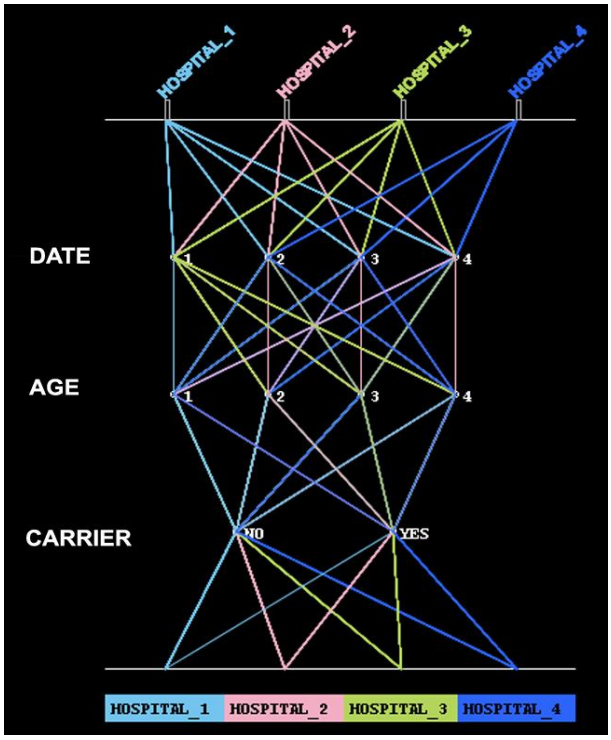

Figure 7: BIOMET3.DAT, same as Figure 3 but with no intermediate axes.

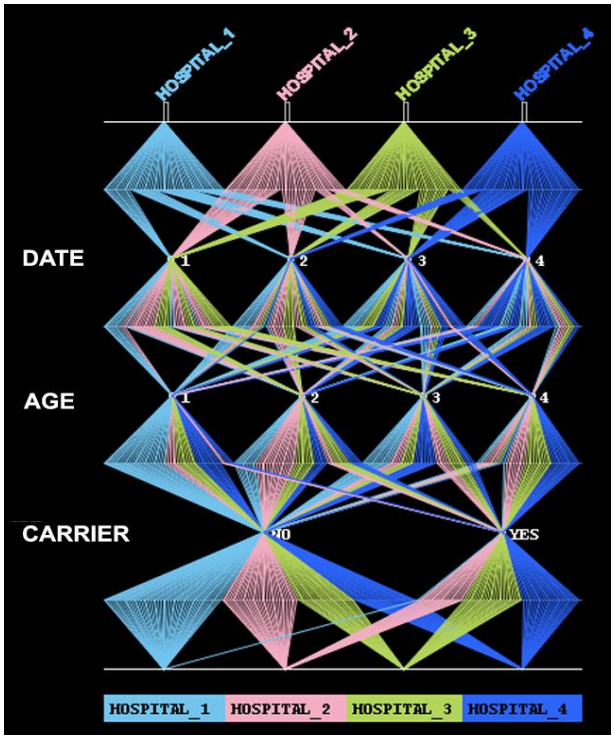

Figure 8: BIOMET3.DAT, same as Figure 3 but with one intermediate axis.

that are less symmetrical can reveal outlier patterns that might go unnoticed in the spread line view. For example, by noting that all hospitals except *Hospital 4* have four tracks leaving their node or, alternatively, all date groups except *1* have four incoming tracks. This reveals an important irregularity – none of the *Hospital 4* patients belong to the *Date 1* group.

In the case of a single intermediate line, lines leaving every category point are spread on the intermediate line to form a cone as in Figure 8. The cone between a category point and its intermediate axis exposes the individual lines that can now be brushed or selected. The cone width at the intermediate axis indicates the cardinality of the category above. The lines within the cone are ordered according to the order of their categories on the next axis. All the lines in a cone going to the leftmost group on the next axis are drawn to the left side of the cone. This avoids occlusion and spatially consolidates similar lines, making the collective patterns easier to recognize. In Figure 8, the lines leading from the *Date 1* group to the intermediate axis show that the *Date 1* group has patients from Hospitals 1-3 (no dark blue lines); almost half of the Hospital 4 patients fell sick in the *Date 4* group. Lines between an intermediate axis and the next categorical axis are somewhat occluded and difficult to follow.

Adding a second intermediate axis between categorical axes adds inverted cones defined by the lines between the second intermediate axes and the next axis as shown in Figure 3. In the inverted cones, lines are ordered according to the order of their groups on the preceding axes. A symmetrical cone is created atop and at the bottom of each cluster point. All line crossing occurs in the area between the two intermediate axes. The ordering of lines in the cones consolidates lines, making the common transitions more noticeable.

4.3. Line Drawing Algorithm

Juxter's line drawing is accomplished by a simple variant of a sorting algorithm, where the sorting takes place at two levels. The algorithm is itself a two-phase algorithm. In the first phase, data structures (mainly Java HashMaps) are created to store references to each of the entities belonging to each category for each axis. These data structures also indirectly store the counts of the entities per category and categories per axis.

The second phase of the algorithm sorts the entities per category and calculates each entity's line coordinates for each axis. The lines coming into a particular category are sorted based on the entities' previous category assignments. The

lines going outward are sorted based on the entities' subsequent category assignments. The main purpose of this two-level sorting is to consolidate the lines of entities making the same transitions across axes. A category's incoming lines form an upside-down triangle of sorted lines. The outgoing lines form another triangle of sorted lines. The length of each triangle's base indicates the relative size of the group of lines; the line colors and order reflect the distribution of the lines across the color encoding dimension (the bottom axis by default) and the adjacent axis. A time complexity of $O(m^2 n \log n)$ is associated with our algorithm, where m is the number of clustering axes and n is the total items.

4.4. Interactions and Links in the Prototype

The possible user interactions are a defining characteristic of any visualization. Juxter supports both basic and complex interactions that allow the user to drive the visualization. The interactions express dynamic operations for incremental, reversible, and quick responses from the visualization [16]. These operations support filtering unwanted items such as lines, groups, and axes; changes in the level of detail; or alternate linked views.

Selection: Individual entities are selected by positioning the tip of the cursor over the entity's line and clicking the mouse. Selected lines are drawn normally (in color), while non-selected lines are drawn in gray. Categories are selected by clicking over the category icon; the result is the same as selecting all entities in the category. Multiple selections can also be defined by dragging the cursor to define an area in the graphic; all lines that pass through the area are selected. The selections can be toggled on and off by depressing the Ctrl key while clicking on a line or category icon. Selected elements are de-selected; unselected items become selected. Selection by toggle allows the accumulation of multiple selections.

Hide/Unhide: Users can control the number of visible axes by hiding and un-hiding them through a menu or GUI.

Reorder: Users can dynamically manipulate the order of individual categories along an axis by drag and drop. The axes order can be specified in a dialog box; re-ordering is immediate. At this time, the top and bottom axes cannot be changed.

Arrow Key Navigation: The arrow keys allow a user to navigate quickly between categories along an axis or between axes. Category selections update immediately, resulting in a scan capability.

Intermediate Axes: Juxter can be dynamically configured to display no (Figure 7), one (Figure 8) or two (Figures 3-6) intermediate axes. This will be discussed in more detail below.

Brushing: Brushing (moving the mouse) over lines or category icons dynamically triggers the display of information about the items in an area below the graphic. The information includes names and counts of items under the cursor.

Linking: Juxter is loosely coupled with a spreadsheet view. The spreadsheet provides detailed tabular information. It also manages the configuration of Juxter. Users can reorder axes by rearranging columns, inserting a column, and pasting categorical information. Juxter immediately updates accordingly. The spreadsheet provides a number of standard operations such as sorting. Juxter and the spreadsheet have a single, shared selection model.

5. DISCUSSION

We summarize below the advantages and limitations of this technique. We also discuss briefly the use of our prototype by a research group of biologists.

5.1. Advantages of the Approach

Categorical data. The most important advantages stem from the choice to visualize categorical data. Categorical data are indifferent to 1) domain-specific details and 2) the diversity of numeric data. The first allows the juxtaposition of categorical data from any number of sources that apply to the same data item set. Many domains may benefit from the ability to integrate data in this way. In systems biology, items such as genes and proteins may be measured and analyzed by a variety of techniques. Comparing these diverse data in their raw state and even in higher-order, derived states is a huge problem, which is our second point. The gene expression data in Figure 1 were derived from reverse-transcription polymerase chain reaction (RT-PCR). Data from different platforms might not be directly comparable without transforming the data into categorical information, for instance by clustering or by binning numeric values. We have had interesting results with our visualization just by categorizing numeric gene expression data over time as either *up*, *down*, or *other* at the various time points.

Multiple dimensional view. The line-spreading technique allows us to visualize many categorical dimensions at once as parallel coordinates. The ability to view many dimensions of data items simultaneously is an important alternative to viewing a series of two-dimensional plots or a single-dimension-reduced view. Viewing the data across multiple dimensions at the same time can provide valuable information about:

- individual entities by the categories they belong to and other entities with similar category memberships
- categories by their number of entities and the numbers other categories on the same axis
- categorizations by the patterns through their categories and by their correlation to other categorizations.

This information can be enlightening and may not be evident in other visualization approaches. The visualization itself with the powerful query and link capability provides both summary and detail information about the data items and collection.

Visual impact. The visual impact of this technique is reminiscent of the ThemeRiver visualization [18], which embodied a number of Gestalt laws [14]. In Juxter, the similarity, continuity, and almost-symmetry of the correlated lines promote pattern recognition. Juxter's line sorting ensures spatial proximity and similarity. The line spreading supports continuity. Matching the top and bottom, either by using the same dimension twice or two dimensions where one is a further partitioning of the other, provides a degree of symmetry that aids in pattern recognition as well as stability to help the user maintain orientation while exploring and reordering the axes and categories.

Simplicity. This visualization seems to be fairly intuitive to users. Juxter accepts simple, tab-delimited text files, making it easy to load data sets. The visualization is readily interpreted, explored, and adjusted on demand. The line-spreading algorithm simplifies pattern identification as well as outlier identification. While Juxter supports confirmatory analysis, its strength is hypothesis generation and exploratory analysis.

5.2. Limitations

Scale. The primary limitation of our approach is the amount of data that can be handled. The visualization works well for data sets of 100 – 400 data items with 5-6 axes and 4-5 categories per axis. Too many categories along an axis make the labels hard to distinguish and, worse, increase the complexity and reduce the gestalt of the graphic. Similarly, large differences in the cardinality of categories along an axis interfere with the gestalt. Finally, a large number of data items stretch the graphic and again ruin the gestalt. When a transition between axes forces lines to be more horizontal than vertical, the patterns in the graphic become very difficult to identify. Allowing the user to hide axes ameliorates this slightly. The excellent method presented in Fua [19] for handling large data sets does not extend to discrete multivariate data though this may be possible using methods in Rosario [7].

Partitioning. Our approach requires a hard partitioning of the data item set across each dimension. That is, each data item must belong to one and only one category along any dimension. Not all data sets are quite so clean. For instance, the correct category for a data item along an axis may be unknown or missing. Or, just as problematic, a data item may fit in more than one category. As mentioned above, too many partitions as well as partitions that vary widely in cardinality also are problematic. These later conditions not only interfere with the gestalt but also severely reduce the amount of information along the affected axis.

The Juxter Prototype. As implemented, Juxter requires a degree of symmetry between the top and bottom dimensions. Either the axes are duplicates or the top is a further partitioning of the bottom partition. Although we have pointed out that this may in fact be fortuitous, we would like to remove this restriction.

Labeling. The labeling leaves much to be desired, especially for labels with complex and/or lengthy text. Some of the interactions are less than optimally tuned. We regret this of course, but our focus has been on exploring the efficacy of the technique.

5.3. Observations

We chose initially to ignore concerns about "optimal" layout. As we gained experience with the prototype and experimented with a variety of data, we began to appreciate the difficulty of even defining an optimal layout. If we look at the parallel coordinates plot as simply a graph, minimizing the number of line crossings might make a cleaner, more interpretable figure [20]. We observe that an optimal layout would not ensure an *ideal* visualization. It is likely more important to lay out the dimensions and categories in a way that makes sense to users to reduce their cognitive load. However, the ideal order is extremely difficult to predict from the labels themselves. Some nominal labels are explicitly ordinal (*first, second*) while others might by ordinal based only on user preference (*green* on the left, *red* on the right, *yellow* in the middle). In other cases, the nominal labels are integer numbers implying order where none exists, as with

index numbers for cluster labels. We believe that the best solution is to allow users to easily re-arrange axes and categories on demand. We also believe that the visual impact of the line drawing algorithm and the filtering capability overcome a degree of imperfection in the category layout.

We found that the vertical stacking of axes provided more space for the labels. At the top and bottom, the labels can be set at an angle to increase the space for long labels, reducing occlusion. This vertical stacking also provides an open space for the category labels where the points of the stacked cones meet.

5.4. Adoption

The prototype was developed to visualize biological data, which are often clustered and annotated. The ability to view biological items such as genes across the derived and annotated information is important to biologists. Recently, scientists at the Pacific Northwest National Laboratory (PNNL) evaluated an interesting metagenomics data set [21] using Juxter. They clearly expressed that the visualization helped them in their analysis. Screen captures of Juxter visualizations of their data were used in an invited presentation and poster at a Principal Investigator meeting for the Department of Energy Genomics: GTL Program [21]. Although the scientists had their data in a huge electronic spreadsheet and had access to powerful bioinformatics tools, this simple visualization quickly provided valuable insight that might have been missed without the ability to look directly at the derived categorical and binned qualitative data in the context of other categorical data. We collaborated to write about the analysis of the metagenomic data using the visualization [15].

5.5. Future Work

Future research will include the development of methods to detect patterns automatically and bring them to the user's attention. We also want to provide additional cues to help the user understand the data and evaluate the patterns. This includes access to mathematical or statistical cues ranging from simple percentages (we already provide counts) to more sophisticated cues to help assess the information content of individual axes and the correlation (or non-correlation) of axes. Our research will extend to larger-scale, linked visualizations. Alternatives include exploring approaches that map discrete to continuous values such as presented in Rosario [7] and Fua [19] or to develop complementary linked visualizations not based on parallel coordinate plots.

Future work also includes enhancements to the prototype such as improving the performance of interactions and calculations, which bog down at the upper limits of our scale, and allowing users to control the color assignment and to dynamically adjust the resolution of hierarchical categorical data.

6. CONCLUSION

This paper presents a technique for visually integrating and exploring diverse information. While the analysis of a single data type is enough to challenge both humans and computers, it is no longer sufficient to analyze one data type in isolation. Users want to explore and understand their data in the context of, or integrated into, related data seamlessly and with minimal constraints.

Our technique visualizes categorical data associated with a data set. This might be meta-data, curated annotations, and attributes as well as the results of other analytical techniques. Cluster membership and uncertainty bins are examples of analysis results turned categorical data. Similarly, any partitioning of the data defines a set of categories; each partitioning is considered a dimension. Our technique allows users to track individual data items across the multiple dimensions and identify items with similar patterns. It allows users to observe the global patterns of items belonging to the same partition, for example a cluster, to characterize that cluster in light of the other dimensions. The visualization is an adaptation of parallel coordinates with a line-spreading algorithm that perceptually organizes the visual items by Gestalt laws to ease the cognitive load for users. We illustrate this technique through a prototype, Juxter, which allows users to dynamically explore a data set.

7. ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Energy (DOE) through the Biomolecular Systems Laboratory Directed Research and Development program at the Pacific Northwest National Laboratory (PNNL). PNNL is a multi-program national laboratory operated by Battelle Memorial Institute for DOE under contract DE-AC06-76L01830.

REFERENCES

[1] D. B. Carr, R. Somogyi, and G. Michaels, "Templates for looking at gene expression clustering.," Statistical Computing and Graphics Newsletter, vol. 8, pp. 20-29, 1997.

[2] G. S. Michaels, Carr, D.B., Askenazi, M., Fuhrman, S., Wen, X., Somogyi, R., "Cluster Analysis and Data Visualization of Large-scale Gene Expression Data," Pacific Symposium on Biocomputing, vol. 3, pp. 42-53, 1998.

[3] A. Inselberg, "The plane with parallel coordinates.," The Visual Computer, vol. 1, pp. 69-91, 1985.

[4] A. Inselberg and B. Dimsdale., "Parallel coordinates: A tool for visualizing multidimensional geometry..” In Proceedings of IEEE Visualization Conference, 1990.

[5] E. Wegman, "Hyperdimensional data analysis using parallel coordinates," Journal of the American Statistical Association, vol. 441, pp. 664-675, 1990.

[6] G. E. Rosario, E. A. Rundensteiner, D. C. Brown, and M. O. Ward, "Mapping nominal values to numbers for effective visualization.” In Proceedings of IEEE Symposium on Information Visualization (INFOVIS 2003), Seattle, WA, 2003.

[7] G. E. Rosario, E. A. Rundensteiner, D. C. Brown, M. O. Ward, and S. Huang, "Mapping nominal values to numbers for effective visualization," Information Visualization; Special issue of selected and extended InfoVis 03 papers, vol. 3, pp. 80-95, 2004.

[8] M. Graham and J. Kennedy, "Using Curves to Enhance Parallel Coordinate Visualizations.” In Proceedings of International Conference on Information Visualization (IV'03), London, England, 2003.

[9] M. Friendly, "Graphical Methods for Categorical Data.” In Proceedings of SAS Users Group International (SUGI 25), 1992.

[10] M. Friendly, *Visualizating Categorical Data*: SAS Publishing, 2000.

[11] M. Friendly, "Visualizing Categorical Data: Data, Stories, and Pictures.” In Proceedings of Twenty-Fifth Annual SAS Users Group International Conference, Indianapolis, 2000.

[12] E. Kolatch and B. Weinstein, "CatTrees: dynamic visualization of categorical data using treemaps," http://www.cs.umd.edu/class/spring2001/cmsc838/Project/Kolatch_Weinstein., 2001.

[13] B. Johnson and B. Shneiderman, "Tree-maps: a space-filling approach to the visualization of hierarchical information structures.” In Proceedings of IEEE Conference on Visualization 1991, San Diego, CA, 1991.

[14] C. Ware, *Information Visualization: Perception for Design*. San Francisco, California: Morgan Kaufmann, 2000.

[15] S. L. Havre, B.-J. Webb-Robertson, B. Gopalan, A. Shah, and F. J. Brockman, "Bioinformatic Insights from Metagenomics through Visualization." In Proceedings of IEEE Computational Systems Bioinformatics (CSB '05), pp. 341-350, Stanford, CA USA, August 2005.

[16] B. Shneiderman, "Dynamic Queries for Visual Information Seeking," *IEEE Software*, vol. 11, pp. 70-77, 1994.

[17] J. de Leeuw, "Here's Looking at Multivariables," in *Visualization of Categorical Data*, J. Blasius and M. Greenacre, Eds.: Academic Press, 1998, pp. 1-11.

[18] S. Havre, E. Hetzler, P. Whitney, and L. Nowell, "ThemeRiver: Visualizing Thematic Changes in Large Document Collections," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, pp. 9 - 20, 2002.

19] Y. Fua,, Ward, M. O., and Rundensteiner, E. A. 1999. "Hierarchical parallel coordinates for exploration of large datasets". In *Proceedings of the Conference on Visualization '99: Celebrating Ten Years* (San Francisco, California, United States). IEEE Visualization. IEEE Computer Society Press, Los Alamitos, CA, pp 43-50.

[20] H. C. Purchase, "Which aesthetic has the greatest effect on human understanding.," in Graph Drawing 97, vol. 1353, Lecture Notes in Computer Science, G. Battista, Ed.: Springer Verlag, 1997, pp. 284-290.

[21] F. Brockman, N. Maltsev, T. Bompada, B. Gopalan, S.M. Li, W. Zhang, J. C. Detter, P. Richardson, and M. Romine, "Metagenome Analysis of Contaminated Sediments at the DOE Hanford Site," in *2005 Genomics:GTL Workshop*. http://www.orau.gov/gtl2005/abstracts/Brockman.pdf, 2005.