VisuMap version 5.0

A High Dimensional Data Explorer

Introduction

The development of information technology made available of new data in unprecedented speed. The ability to quickly explorer data becomes a key for the success in businesses, industry and scientific research. At the same time, the advancement of machine learning in the last twenty years yielded tremendous numbers of software and hardware tools to address the new challenge in data exploration.

VisuMap as a software application aims to provide a tool with the latest software and hardware technologies for the visually exploration of high dimensional complex data. VisuMap unleashes the perception power of experienced human eyes; and enables people to quickly analyze tables with large number of rows and columns from different perspectives.

The Challenge

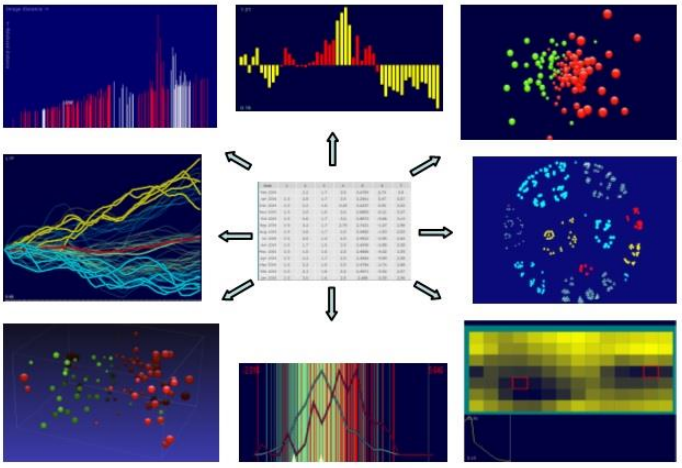
A major challenge for visualization method is the high dimensional nature of many data tables: If a table contains more than 3 data columns a visualization method must convert the high dimensional data to 2- dimensional graphics with certain means.

Most available visualization software relies on the user to select few data columns or calculate few attributes for the visualization purpose. Those software applications share the common limitation: they all require good knowledge about the data.

Some software alleviates the limitation with interactive exploration support; and some others resort to more generic multidimensional scaling methods. Yet, there is no software on the market that treats the *dimensionality reduction* as a central task. Thus, exploratory data analysis of true high dimensional data remained so fare an ad-hoc and difficult undertaking.

The Solution

With a *visual centric design* VisuMap is a software application developed from ground up to target the analysis of high dimensional data. After data has been imported into the application, data points will be represented by dots, glyphs, curves or bars in various *visual data views*. Each of those data view offers a different perspective of the data. Those views are linked with each other; so that a user can simultaneously investigate a dataset from different perspectives with multiple views.



At its core VisuMap implements a collection of *dimensionality reduction* algorithms which map high dimensional data to 2D or 3D maps. Tables with large number of columns of numerical and textual data can be quickly transformed into maps which give the user an overview of the whole dataset.

On top of those core services VisuMap offers whole palette of features to support interactive visual data analysis. Those features include, among the others, data clustering, annotation, drilling, sorting, linking and 3D animation.

For advanced users VisuMap offers a script interface to automate frequently performed tasks, as well as a library plug-in interface to implement domain specific extensions.

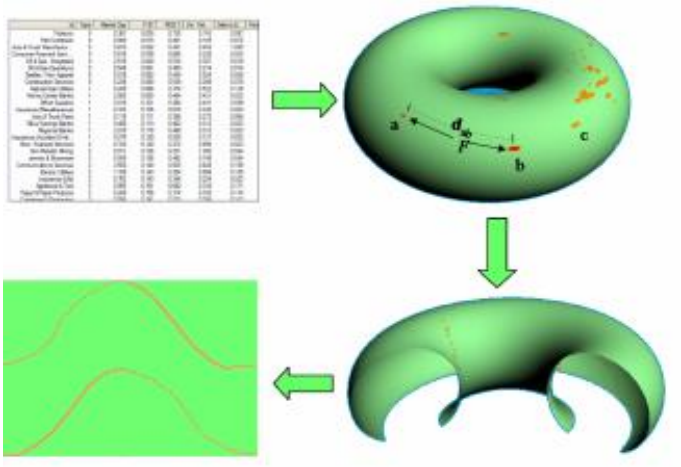
Main Services of VisuMap

Dimensionality Reduction

A table in a relational database can be considered as a high dimensional dataset by viewing each data record as a data point in a high dimensional space whose dimension is the number of columns of the table. One of the core services of VisuMap is to *map* the high dimensional datasets to 2 or 3 dimensional spaces while preserving relevant information.

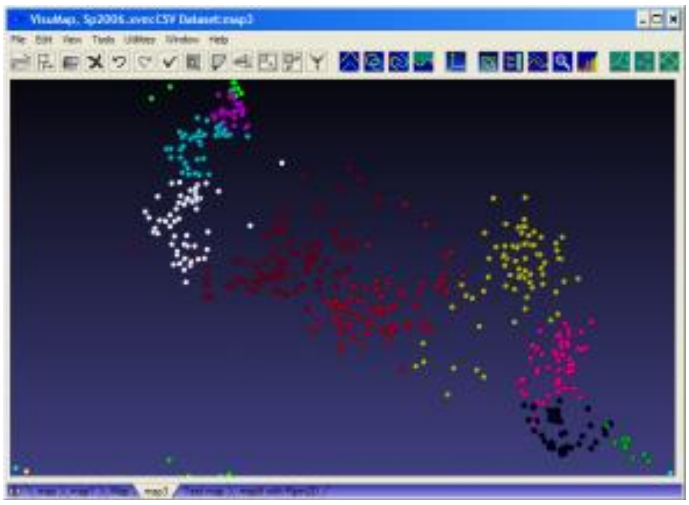
Since different applications require different information set, it is critical to provide different methods to preserve different type of information. Whenever available, VisuMap always resort to GPU hardware to achieve the best performance. VisuMap implements the following dimensionality reductions algorithms:

- **Principal Components Analysis** (PCA). A linear projection algorithm. This method offers a fast high-level overview of a dataset; it is good for simple mainly linear datasets.
- **t-SNE Embedding**: A non-linear dimensionality reduction algorithm. It has been widely used in machine learning to visualize cluster structures among large high dimensional data sets.
- **Sammon Mapping**: A widely used non-linear mapping algorithm. This method offers a good overview of the global characteristics of the dataset while provide only limited capability to preserve non-linear features.
- **Curvilinear Component Analysis** (CCA). A non-linear mapping algorithm with time dependent strategy. This method preserves local non-linear features much better than the previous methods. This method is relatively more calculation intensive.
- **Relational Perspective Map** (RPM). A non-linear mapping algorithm that is based on simulation of dynamic system on closed manifolds. This method performs global segmentation of the dataset and present local features in a non-overlapping manner. This method is relative calculation intensive.



RPM Algorithm

VisuMap implements all algorithms in a consistent way, so that the user can easily generate different maps from the same dataset and quickly compare them from different perspectives. VisuMap user interface is designed like a modern spreadsheet application so that it is easy to use to for people with experiences with spreadsheet applications. The following picture shows a snapshot of VisuMap user interface:



Metric and Filter

VisuMap implements the concept *metric* and *filter* to offer maximal flexibility in defining information set on datasets. Whereas the dimensionality algorithms focus on mapping high dimensional data to low dimensional space, the metric and filter mechanisms enables the

user to directly specify different information set on the same dataset.

VisuMap currently supports the following metric:

- Euclidean distance.
- Mahalanobis distance.
- Symmetrical information.
- Hamming distance.
- Direct distance matrix.
- Pearson correlation
- Speaman ranking correlation.
- Kendall ranking correlation.
- Graph tree distance.
- Intersection distance
- Wave hedge distance
- Arbitrary custom metric implemented through the VisuMap plug-in interface.

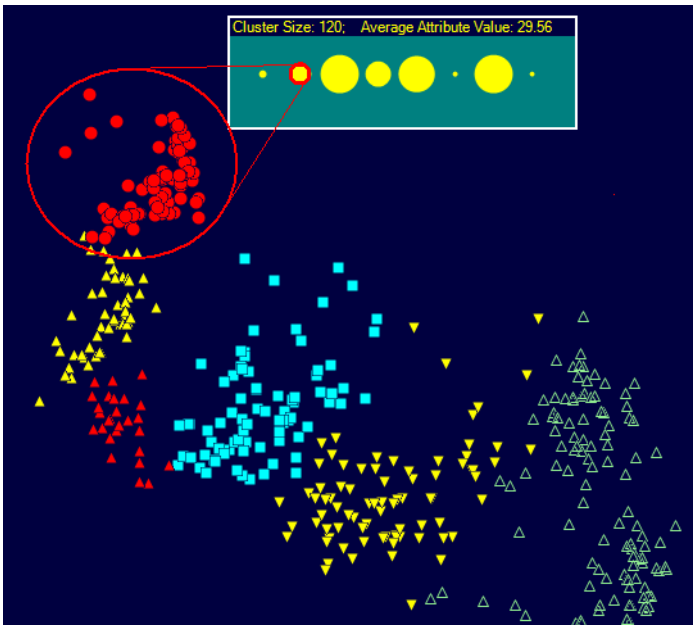
The filter concept allows users to quickly enable, disable or scale data columns. Filters in VisuMap are implemented as stand-alone objects, so that users can share and copy filters between datasets.

Data Clustering

VisuMap implements a collection of clustering algorithms with which the user can generate clusters and annotate their representation with different colors. The clustering algorithm also allows user to reduce dataset size by generate new data points from clusters. VisuMap can to cluster dataset with millions of data points, and reduce datasets to appropriate size for interactive investigation.

Currently VisuMap supports the following clustering algorithms:

- **K-Mean Clustering**: a classic clustering algorithm for multidimensional dataset.
- **Agglomerative Clustering**: an algorithm to cluster datasets with any well-defined distance matrix.
- **Self-Organizing Map** (SOM): a clustering algorithm that preserves topological properties underlying a dataset. The self-organizing map is especially appropriate to reduce size of large dataset.
- **Self-Organizing Graph** (SOG): a proprietary extension of the SOM algorithm that performs data clustering according to arbitrarily structured network.
- **Metric Sampling**: a special clustering algorithm for non-Euclidian datasets, e.g., datasets in form of tree or network structures.
- **DBSCAN/HDBSCAN**: A widely used clustering algorithm for relatively low dimensional data.
- **Affinity Propagation**: a recent clustering algorithm for generic data equipped with similarity matrix.



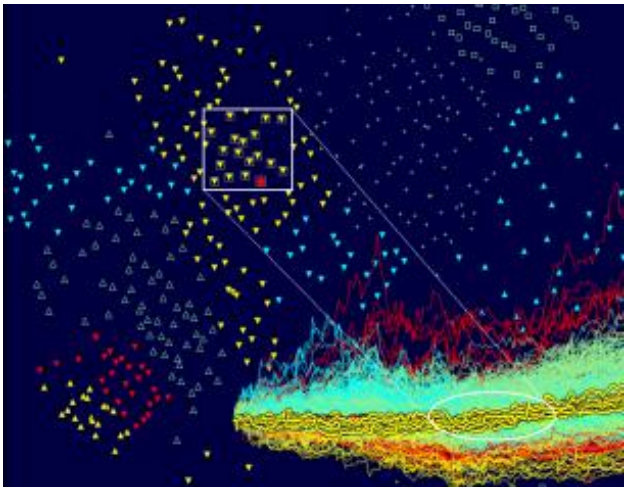
Clustering and Coloring dataset
with K-Mean Clustering Algorithm

Interactively Linked Views

With VisuMap users can explore a dataset from different perspectives simultaneously with multiple views. The data points are represented as glyphs, curves or bars in different views. Changes made in one view will be immediately reflected in other views.

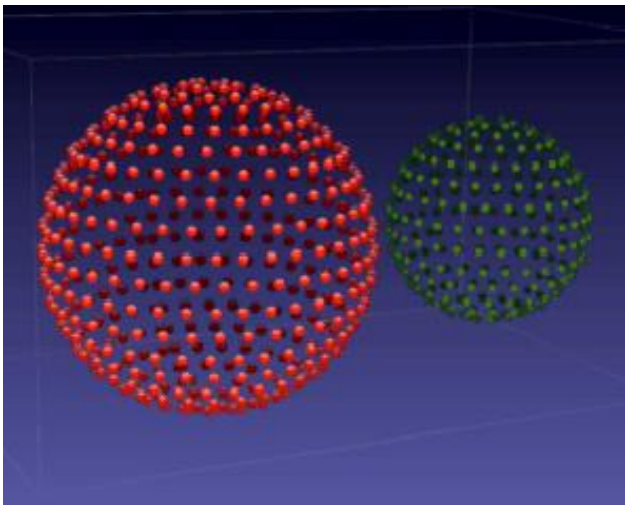
The following picture shows a dataset with two views — the RPM map view in which each data point is represented as a colored dot; and the

value curve view in which each data point is represented as colored curve. When the user selects a subset of data point in the RPM map, their corresponding curves will immediately be highlighted in the curve view.



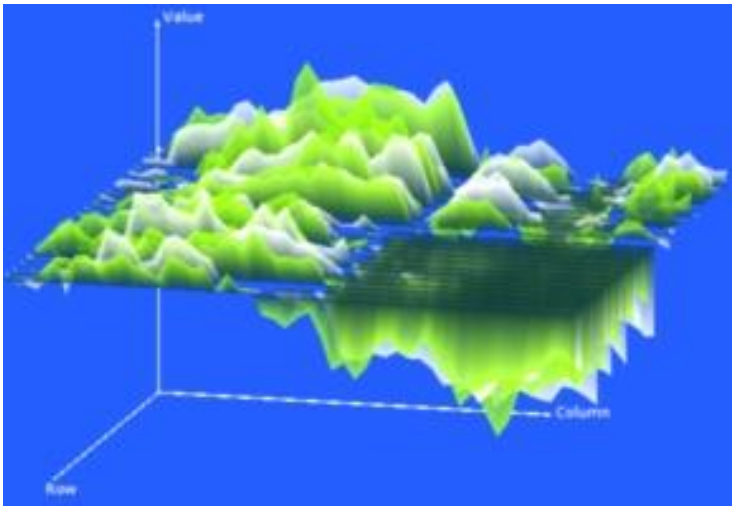
VisuMap implements the following views:

- The scatter plot view in which a data point is represented as a glyph. The distance between the glyphs will be generated by one of the dimensionality reduction algorithms.
- The curve plot view in which each data point is represented as curve.
- The details view that represents the whole dataset as a table with one row for each data point showing the details of the data point. Special user interface optimization has been implemented so that the user can directly load tables with up to 100 000 columns into the application.
- The table view that presents the complete dataset as an editable table.
- Shepard diagram that plots the relational distance in high dimensional space against the image distance in 2D/3D space.
- 3D animation view that enables users to navigate in a 3D space. The 3D animation view provides fast animation by taking advantage of graphics card.



3D animation view

- Attribute map that shows the data attributes (e.g., data columns) as scatter plots. The user can then study the relationship between the attributes and perform attribute related operations.
- Spectrum view that displays data of a selected dimension as spectrum style map.
- Bar chart view that displays data of a selected dimension as a bar chart map.
- Mountain view that displays a complete data table in a mountain landscape 3D style.



Mountain View

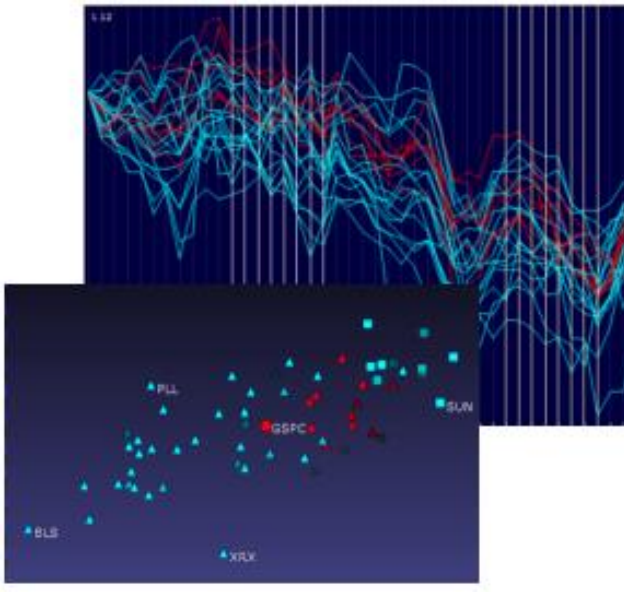
- Atlas view that enables user to organizer and compose visual information of different types in a single window.

Dual Mode Data Drilling

To explorer complex large data sets it is prevalent to be able to focus the investigation on subsets of data. VisuMap provides more than a dozen of methods to enable users to select data interactively and visually. Users can then successively drill into data subsets and explore them with any of the available views.

Moreover, most views in VisuMap can operate in either data row-mode or column-mode. In the data row- mode the user drills into the data set by selecting data rows in the dataset table. In the column-mode the user drills into the data set by selecting columns in the dataset table.

The column-mode provides a quick way to explore patterns and regularities underlying subset of attributes. For example, the following curve plot view shows the daily price history of some traded stocks. The diagram indicates two intervals in which the stocks experienced collectively down-turns. If we are interested in correlation between these stocks during these two periods, we can simply select those attributes in column-mode and open the PCA window to study them in 3D view.



Programming Interfaces

VisuMap provides two types of programming interfaces for advanced usages:

- The *script interface* enables users to automate most interactive tasks with script written in standard JavaScript and Python languages. VisuMap offers simple GUI user interface to create, test, execute and integrate scripts. The script interface can also be used to interact with third party applications.
- The *plug-in interface* enables advanced users to extend VisuMap with domain specific features. Plug-in modules are libraries implement in any language supported by the .Net platform. Plugin interface has often been used to customize VisuMap to support domain specific need, like single cell RNA sequence data profiling.

System Requirements

- Windows 10 and above
- Microsoft .Net 4.6.1 Library or higher
- Microsoft DirectX 11 Runtime Library.
- DirectX 11 compatible graphic card.

Conclusion

VisuMap offers a unique and powerful collection of services for visual exploratory analysis of high dimensional dataset. By focusing on the goal of dimensionality reduction and by employing the latest software and hardware technologies VisuMap achieved a flexible, high performance and user-friendly tool for knowledge discovery.

Contacts

James X. Li, Ph.D.,
CEO & Chief Scientist
252 Edgehill Dr. N.W.
Calgary, Alberta T3A2W8

Email: information@visumap.com
http://visumap.com