FULL TITLE
ASP Conference Series, Vol. **VOLUME**, **YEAR OF PUBLICATION**
NAMES OF EDITORS

Interactive Data Visualization and Pattern Discovery with Mirage

Tin Kam Ho
Bell Labs, Lucent Technologies
600 Mountain Ave., 2C-279, Murray Hill, NJ 07974, USA
tkh@research.bell-labs.com

Abstract.
Pattern discovery often occurs in studying correlations of data viewed from multiple perspectives. For high dimensional, low level sensory data such as images, it is especially important that the input can be examined in its raw form, along with features extracted according to the application's concern, plus all intermediate results of manual or automatic analysis. To facilitate this, flexible and effective exploratory tools that can handle diverse data types, a wide range of objectives, large data volumes, and variable demands on speed are in critical need. I describe our experience with Mirage, a research software tool for open-ended interactive pattern discovery. Applications to analyzing photonics simulations and in the Virtual Observatory raised interesting challenges on the tool's organization and the data visualization algorithms. We describe our progress on designing effective modularization and connection interfaces to local or remote analysis libraries, and approaches to highlighting data relationships between high-dimensional spaces or across contexts of different resolutions and data types.

1. Useful Features of Mirage

Analysis of large scale astronomical surveys shares many difficult challenges with other areas of study that involve large imaging datasets, such as Earth observations, security surveillance, and design and monitoring of complex engineering systems. A tool is needed to support pattern discovery across diverse data types while enabling analysis suited to each type. These may include attribute tables of objects or events with both numeric and symbolic fields, time series, images, maps, video sequences, and spectra.

Analysis of such data typically is an extended process that involves sophisticated mathematical or statistical methods and expert judgements. Ideally, the analysis tool should support both human expert centered explorations and automatic pattern discovery algorithms, and allow seamless exchanges between the two sides. This way, automatic algorithms can be guided by domain expertise, and human judgement can be assisted by consistent and efficient automatic processing. The tool also should leverage existing visualization and analysis techniques while providing room for continuous addition of new techniques acquired or invented in the future, and be scalable to handle a large variety of data volumes and processing speeds.

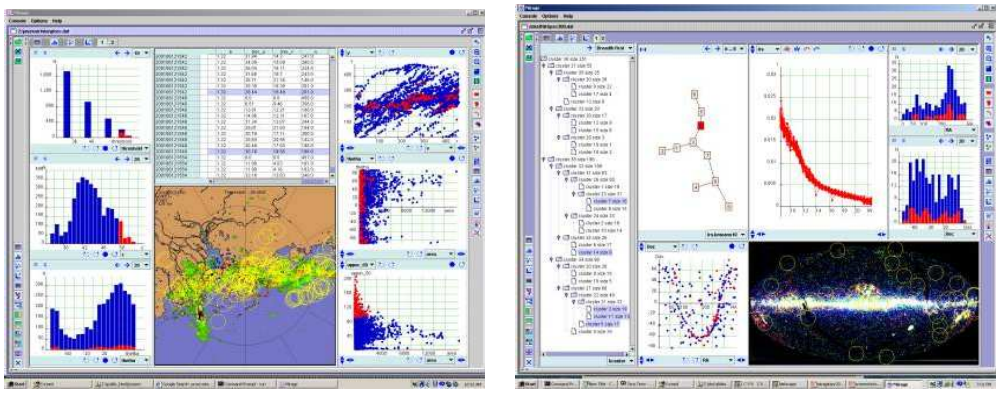


Figure 1. Mirage screen shots: (Left) Linked views of a table, 3 histograms, 3 scatter plots, and a map. A subset of data is selected in one histogram and broadcasted to other views. (Right) Linked views of two cluster trees, 2 histograms, a scatter plot, a spectrum plot, and an image. A step on a leaf-to-leaf walk in a cluster tree is broadcasted to all other views.

We describe **Mirage**, a tool designed to address such needs and challenges (Ho 2002-2004). Many queries about object properties and the relationship between different objects can be translated into graphical operations in Mirage. This is done mainly through a versatile assembly of displays each suited to a specific kind of data structures like vectors, histograms, or graphs. By bringing diverse displays together, raw data can be shown alongside results from different stages of analysis. This can help sanitary checking in data preparation, proposing models and algorithms by an initial exploration, evaluating tentative modeling and predictions, and validating and interpreting final results. We describe below the several key design ideas.

Simultaneous, multiple views in a flexible layout.

For the primary goal of visualizing data geometry, Mirage provides several basic tools of statistical graphics like histograms, scatter plots, and parallel coordinate plots. These are linked to displays for non-numerical data, such as text, images, maps, graphs, and trees. Custom-developed displays for application specific data types, such as video and audio clips, coordinated matrix of similar plots, image arrays with synchronized zooming, can be plugged in via a standard interface. The displays are organized as a set of pages; each page can be arbitrarily tiled and any display can go into any tile by drag-and-drop.

Core exploration commands and display-specific manipulations.

A set of core commands are implemented by each display module. These are for broadcasting the local selection of a set of data points, highlighting or coloring a broadcast selection, deleting the selection, canceling the highlights or colors, and for switching between monochrome and color display modes. In addition, each module can provide local operations suitable for the specific data type, such as manipulation of image color maps, changing resolution of histograms, and zooming in and out of a specific display focus. The selection-broadcasting mechanism is very effective for tracking correlations. Items of interest found

from one view of the data set can be easily traced in other views. Coordinated sequences of selection and broadcasting can be used as packaged commands to track the effects of systematic changes in data attributes. An example is a step-by-step broadcasting of a standard traversal of a cluster tree, or a sequential walk on consecutive bins of a histogram. When other displays highlight the corresponding data items, one can easily find out if the items are also similar in terms of other attributes.

Facilities for examining tentative modeling.

A tentative classification can be represented as one column in the data matrix, and displayed as one attribute. Selection in that display can be broadcasted to other displays where the effects of the classification can be examined. This way one can easily compare alternative classifications to refine the problem formulation, select features for the classifier, and polish the classifier design. Moreover, data can be imported on the fly by adding columns and rows to the data matrix. Columns added can show newly observed or computed attributes, including classification decisions, cluster memberships, or predictions from external algorithms. Rows added can be additional samples from the same source, which can be used to check the correctness of a computed model. Continuous adding and removing rows via a programmatic interface turns the software into a monitoring tool.

Extensible software design.

The software is designed to be extensible in different ways. Specialized displays can be added as plug-in's. Exploratory actions are organized around a small command interpreter, which can be activated via an application interface. In addition, the software has a slot for plugging in an "Action" panel, which can be used to encapsulate data updating and analysis operations useful for the current task. The Mirage window can also be embedded into and invoked by wrapper programs which can implement more sophisticated data access methods and decision support systems. We continue to investigate better ways to cooperate with external analysis code while maintaining a simple and modular core architecture.

Scripting for repeatable, cooperative investigation.

Data analysis is seldom an individual effort. To enable easy repetition and sharing of exploration commands and results by a team of analysts, Mirage provides text-driven commands and scripting facilities. Scripts can be passed around for replay. They can also be systematically constructed by simple programs to make animations.

2. Applications and Continuing Development

Mirage has been posted on the web since late 2002, with enthusiastic responses. Its design has been stimulated by several extended applications in our environment. Mirage was used in studying simulations with complex models of optical physics to yield new designs of optical fibers and identified practical architectures of optical line systems (Ho et al. (2005)). It served to monitor IP traffic in an MPLS-based network management system (Thottan et al. (2003)). It

has been linked to the Virtual Observatory via the IVOA Client package and a sophisticated image display (Carliles (2003)).

A key theme in the design of Mirage is to make data exploration activities easy. Our goal is to allow the scientist to ask pointed questions about their data, such as:

- "I see this point standing out in this plot. Which object is it? Does it stand out in other plots too?"
- "The data seem to form two distinctive groups in this plot. Will I see them in two groups if I do a plot with two other attributes?"
- "Are there any objects from our dataset that happen to locate near the right edge of this image? How does a histogram of their brightness look like?"
- "These cases show up as a cluster in the upper-left corner of this plot. Will I see them exceeding threshold T in the temperature histogram? Do these points distribute evenly over the map?"
- "If I label those two groups of points as two classes, will an automatic classifier be able to predict the right class using this set of attributes? Which attributes are the most critical in the prediction? How likely will it fail if I take away the first two attributes?"

We seek to translate such questions a scientist may ask about his data to intuitive, effortless graphical operations, with immediate, visual answers. We hope to further understand the scientific discovery process and how to facilitate it by computer science methods through a continuous dialog between the domain scientists, statisticians, and the computer scientists.

References

Mirage project site: http://www.cs.bell-labs.com/who/tkh/mirage.

Ho, T.K., Exploratory Analysis of Point Proximity in Subspaces, *Proceedings of the 16th International Conference on Pattern Recognition*, Quebec City, Canada, August 11-15, 2002.

Ho, T.K., Mirage: A Tool for Interactive Pattern Recognition from Multimedia Data, *Proceedings of Astronomical Data Analysis Software & Systems XII*, Baltimore, MD, October 13-16, 2002, 339-342.

Carliles, S., Ho, T.K., O'Mullane, W., VO Enabled Mirage and the IVOA Client Package, *Proceedings of Astronomical Data Analysis Software & Systems XIII*, Strasbourg, France, October 12-15, 2003.

Thottan, M., Swanson, K., Cantone, M., Ho, T.K., Ren, J., Paul, S., SEQUIN: An MPLS Network Monitoring System, *Bell Labs Technical Journal*, **8**, 1, 2003, 95-111.

Ho, T.K., Interactive Tools for Pattern Discovery, *Proceedings of the 17th International Conference on Pattern Recognition*, Cambridge, U.K., August 22-26, 2004, Volume 2, 509-512.

Ho, T.K., Salamon, T., Freund, R., White, C., Hillyer, B., Cowsar, C., Nuzman, C., Kilper, D., Simulation of Power Evolution and Control Dynamics in Optical Transport Systems, *Bell Labs Technical Journal*, **10**, 1, January-March 2005, 119-137.