Information Visualization for the Collaborative Analysis of Complex Data

Paul Craig[1], Néna Roa-Seïler[2], Ana Delia Olvera Cervantes[3], Marco Polo Tello Velasco[4], Martín Reyes García[5]

Abstract

This paper describes state of the art information visualization research undertaken within the Sistemas Interactivos group at the Universidad Technologica de la Mixteca to make large scale complex data-sets more accessible for scientists and the general public. Our central thesis is that by making data more accessible we can improve the efficiency of our scientists and democratize data for the general public toward improving development. This is particularly important in traditionally disadvantaged areas such as the state of Oaxaca where the state of development lags behind that of other parts of Mexico and the efficient use of human and physical resources are essential to improve this situation. We also demonstrate that information visualization is an important tool for collaboration and would like to encourage the use of information visualization in university systems, such as the SUNEO of Oaxaca (Seara Vázquez 2009), where regular face to face collaboration between staff can be difficult due to the large distances between campuses. The examples of information visualization presented in the paper are applied in the fields of bioinformatics, taxonomy, logistics, intelligent agent research, spoken language systems and information retrieval.

Keywords: Information Visualization, Interaction Design, Bioinformatics, Logistics

INTRODUCTION

Information Visualization (Card, Mackinlay et al. 1999) is an area of research concerning the design and development of methods for presenting abstract data using interactive graphics to improve cognition. This is applied in areas where large, complex or diverse

[1] PhD in Information Visualisation, profesor-investigador at the Universidad de la Mixteca e-mail: p.craig@utm.mx.
[2] PhD candidate in Emotional Interaction, profesor-investigador at the Universidad de la Mixteca e-mail: n.roa-seiler@mixteco.utm.mx.
[3] PhD candidate in Statistics, profesor-investigador at the Universidad de la Mixteca e-mail: ana.olvera@mixteco.utm.mx.
[4] PhD in Business, profesor-investigador at the Universidad de la Mixteca e-mail: mptello@mixteco.utm.mx.
[5] Masters student in Business at the Universidad de la Mixteca e-mail: oceano115@hotmail.com.

datasets are common, such as bioinformatics, data analysis, business, criminology etc, with applications normally supporting tasks such as data exploration, browsing, search and analysis. By using a visual representation of the data IV utilizes the ability of humans to extract information efficiently and effectively using the visual analogue. IV avoids the cognitive overhead of interpreting text or numbers to understand a data-set and communicates data in a more intuitive efficient manner. In this paper, we introduce some of the fundamental concepts of Information Visualization and describe our work on the use of visualization, and in particular animated visualization, for complex systems. This includes demonstrations of applications developed by the authors for ecosystems (using taxonomy) (Craig and Kennedy 2008), genetic networks (Craig, Cannon et al. 2013), human emotions (Craig and Roa-Seïler 2013), Mexican history (Craig 2012), dialogue (Craig and Roa-Seïler 2012), logistics and statistics.

Information Visualization.

The classic definition of Information Visualization (Card, Mackinlay et al. 1999) is presented by Card, McKinlay y Shniederman, as;

> "the use of interactive visual representations of abstract data to increase cognition"

Here abstract data generally means data without a concrete physical form 3 and cognition is the process of thinking and reasoning. Information visualization deals with abstract qualities such amounts of money, language, biological measures, votes etc and visualizations are designed to make the data more accessible by helping the user develop cognition by generating and using a mental model 8. There benefits of IV are manifold. In the first instance visual representations of data makes it easier to extract information from data. It is easier to find patterns in data and a visual representation can help comprehension and the retention of information for longer periods of time. IV can also improve the user experience by allowing them to explore the data more effectively and efficiently. This is particularly useful when data analysis involves the repetition of similar tasks or the scale of the data makes a numeric or textual representation difficult to manage. IV can also be used discover unsuspected patterns. This is often the case when the display is used for scientific

data analysis, a classic example of this being the map used by Dr. John Snow in 1854 to link the cholera deaths to the consumption of water from unsanitary pumps (Figure 1).

Figure 1. The map used by Dr. John Snow in 1854 to trace the causes of cholera in central London.

Ultimately, information visualization application can have an economic impact saving time and money in the workplace by allowing us to be more efficient and productive when performing data analysis. IV can also have a considerable social effect by making data more accessible to a greater proportion of the population encouraging them to take a more active role in society.

The first theory of graphic symbols was introduced by cartographer Jacques Bertin in his seminal 1967 thesis Semiologie Graphique (Bertin 1983). In this work, Bertin first presents his theory of marks, visual variables and characteristics. Marks are graphical primitives such as points, lines and areas. Visual variables are graphical properties such as color, position and shape. Features describe the tasks that can be performed using these marks and visual variables. In 1986 the list of visual variables was expanded by Jock D. Mackinlay with visual variables ordered according to their accuracy for various tasks (Mackinlay 1988).

The next important figure in the development of information visualization was the statistican Edward Tufte (Tufte 1983; Tufte 1991). Tufte dedicated himself to a minimalist approach with designers encouraged to minimize the ink to data ratio and avoid unnecessary graphics (which he called 'chartjunk). Tufte also introduced the concept of "lie factor" which is defined as the size of the effect in a graph divided by the size of the effect in the data. This lie factor is evident in Figure 2 with the use of an area to demonstrate uni-dimensional data. This chart also demonstrates ‘chartjunk’ with the image of the doctor repeated unnecessarily.

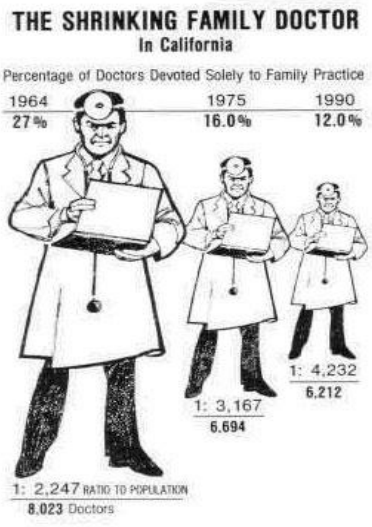

Figure 2: Demonstrating ‘chartjunk’ and the ‘lie factor’ as defined by Tufte.

An important group of guidelines for information visualization are those presented by Ben Shneiderman in his 1996 article, 'The eyes have it' (Shneiderman 1996). This article introduces Shneiderman's mantra of information visualization "overviews first, zoom and filter, then details requested '. The article also introduces us to the 'taxonomy of data by type of work' defining seven types of tasks and seven types of data as a guide to classify the techniques and visualization applications. This taxonomy allows developers to find and re-use existing IV applications in diverse application areas with similarly characterized data. This means having, for example, an employment hierarchy displayed in a similar way to a taxonomic or file-system hierarchy.

Animated Information Visualization.

Previous work in the area of animation for information visualization (Wertheimer 1961; Ware, Bonner et al. 1992; Bryant, Milosavljevic et al. 1998; Holmes 2005) has focused on issues such as the use of animation as another dimension for data visualization, the expressive qualities of motion (Bartram 1997), movement to communicate the structure of objects for representations in three dimensions (Bederso, Meyer et al. 2000; van Wijk and Nuij 2003), animation to smooth the transition between different views of the data (Mackinlay, Robertson et al. 1991; Robertson, Mackinlay et al. 1991; Yee, Fisher et al. 2001; Fekete and Plaisant 2002) and animation to show changes over time in spatial data (i.e. animated maps) (Card, Robertson et al. 1991; Andrews 1995; Ware and Franck 1996). Our own contribution in this is the use of animation to show changes over time for abstract data (Craig, Kennedy et al. 2005).

Motion is a very powerful and expressive visual variable and expressive but it can also give rise to problems and it can be difficult to implement animation in an IV display. In the first instance, motion can be distracting and interrupt the communication of the information carried by other visual variables. There is also the problem that motion is temporary and it's easy to forget what has been seen if a view changes. In order to overcome these limitations we have developed a set of guidelines for the proper use of animation to display information in IV interfaces (Craig, Kennedy et al. 2005). These are as follows;

1. Animated views should be configured so that the motion has relevant meaning.
2. The pace and direction of the animation should be controlled by the user.
3. The motion of objects should be smoothed and regulated to avoid the undesirable effects of erratic or unpredictable motion (e.g. interpolation across time and distortion of the display space).
4. Static views (coordinated and linked) should be available to help the user read a pattern once it is detected.
5. Static views (coordinated and linked) can be used to help the user interpret the animation.

6. The animated view can be used as a static view (see 4 and 5) when the animation is paused.

These rules have been successfully applied by the authors for the development of a number of information visualization applications.

METHODOLOGY

Each of the projects described in this paper has been realized by developing a close working relationship with a sample of our target users and making use of expert knowledge throughout an iterative software life-cycle model with successive prototypes being developed to add new functionality and refine existing features to improve usability. Each project starts with an initial requirements analysis phase to ensure that the user's requirements are properly understood and are best served by the development of an IV application. Normally, information visualization can help a scientist if the data is large scale or complex and the objectives are not well defined from moment to moment. This will be the case if the scientist's want to explore their data, find unsuspected patterns or the results or the path to the results are not straight-forward. After the initial requirements analysis we develop low level prototypes (prototypes or storyboards) to test ideas for potential solutions before proceeding to actual software prototypes. Software prototypes normally pass through three or four cycles of development broken by evaluations and further requirements analysis. The following section describes some late-stage software prototypes.

RESULTS

Information visualization applications have been developed to help scientists annotate taxonomic data, explore dialogue data, analyze human emotions, explore Mexican history and find patterns in microarray time-series data.

The Concept Relationship Editor

The Concept Relation Editor (Craig and Kennedy 2008), shown in Figure 3, is an application designed to allow biologists to explore specify, explore and edit relationships

between taxonomic concepts. This design uses an interactive space-filling adjacency layout (Stasko, Guzdial et al. 1999) optimized for readability. The interface shows two side-by side classifications. Users can explore by clicking on taxa names to change the focus of a classification. Here views are distorted and animation is used to smooth the transition. Relationships are specified by dragging between names to draw a line.

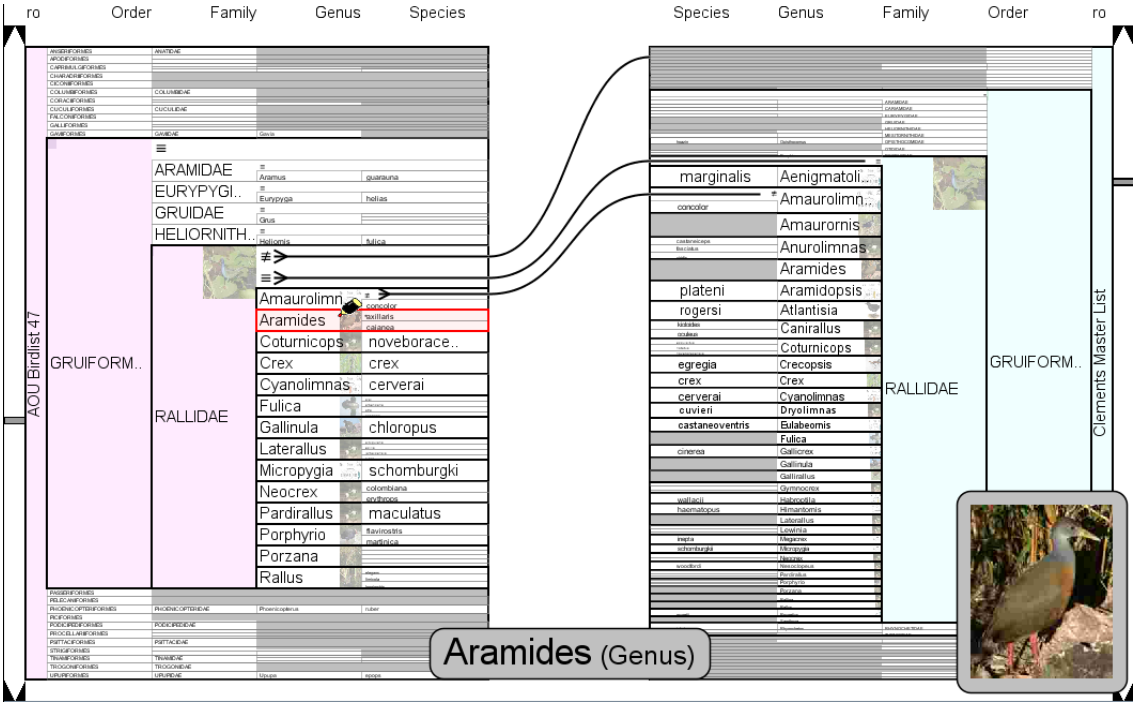

Figure 3. The Concept Relationship Editor

The Dialogue Explorer

The Dialogue Explorer (Craig and Roa-Seïler 2012) is a novel vertical timeline information-visualization technique developed to support the analysis of human-computer dialogue data. The technique uses combined linked views including distorted views to effectively communicate the timing of dialogue events while presenting text in such a manner that it is easily readable. The application also demonstrates the use of the use of animation to see changes over time in abstract data. When the user presses the 'play' can hear the audio of the dialogue while visualization is animated to reflect the change in temporal context.

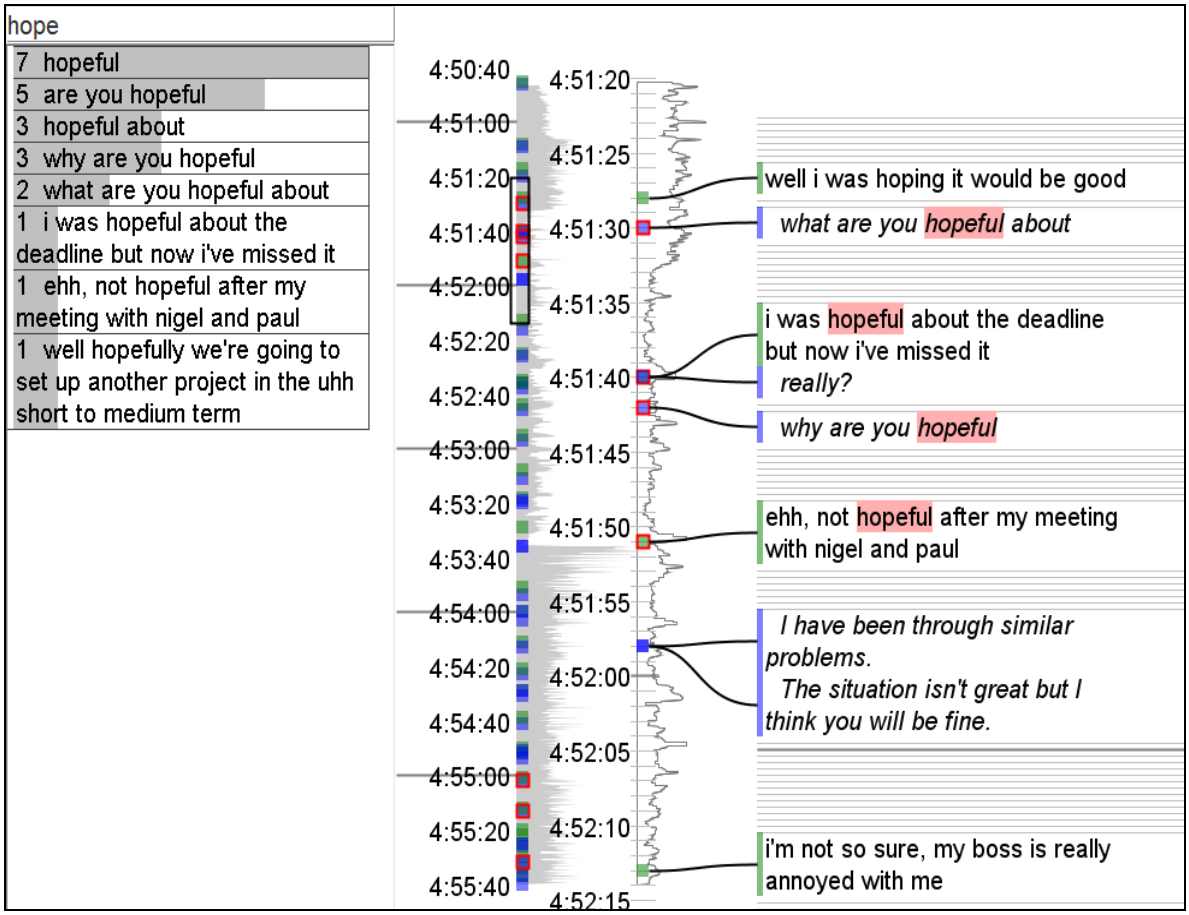

Figure 4. The Dialogue Explorer

The Cluster Explorer

The Cluster Explorer (Craig and Roa-Seïler 2013) combines multidimensional scaling and hierarchical clustering to support the exploratory analysis of multidimensional data. The technique displays the results of multidimensional scaling using a scatter plot where the closeness of any two items' representation's are approximate to their similarity according to a Euclidean distance metric. The results of hierarchical clustering are overlaid onto this view by drawing smoothed outlines around each nested cluster. The difference in similarity between successive cluster combinations is used to colour code clusters and make stronger natural clusters more prominent in the display. When a cluster or group of items is selected, multidimensional scaling and hierarchical clustering are re-applied to a filtered subset of the data, and animation is used to smooth the transition between successive filtered views.

As a case study we demonstrate the technique being used to analyze survey data relating to the appropriateness of different phrases to different emotionally charged situations.

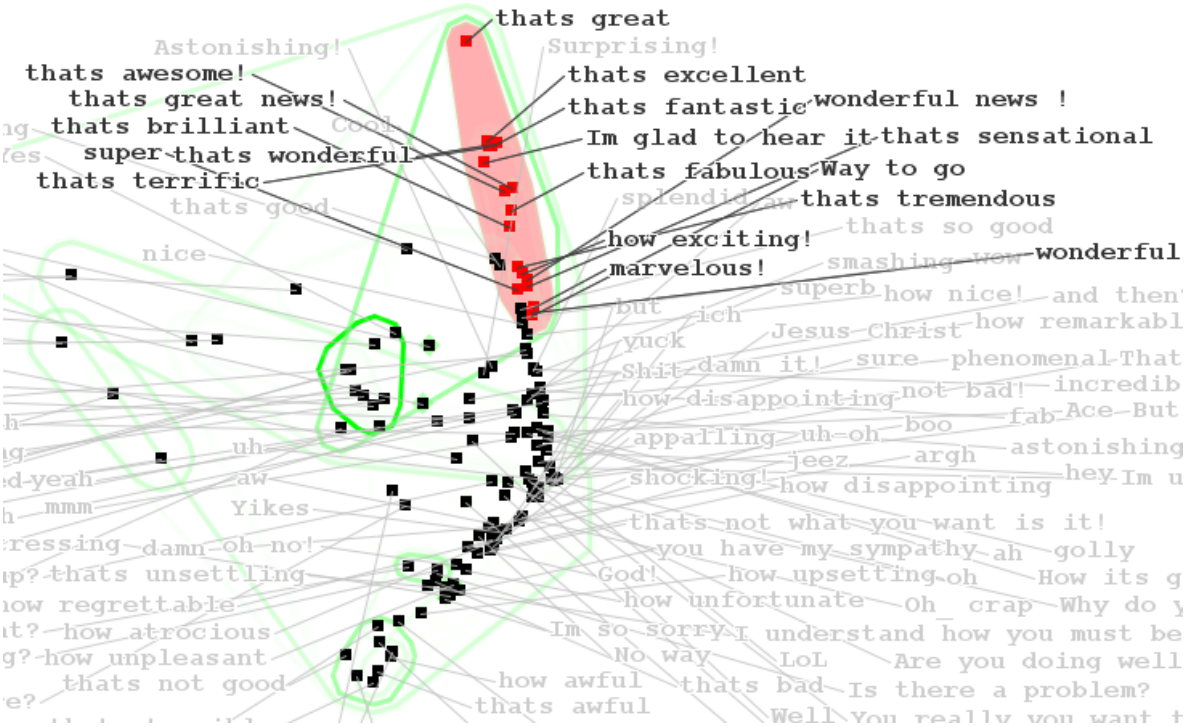

Figure 5. The Cluster Explorer

The Mexican History Browser

The Mexican History Browser (Craig 2012), shown in Figure 6, is easily navigable animated historical map for information retrieval. In this application clustering over time, space and category is used to group events. This helps us to simplify the view and organize data for navigation. It also allows us to describe events using lists rather than disorganized clusters of labels. The application was evaluated and shown to improve user satisfaction and efficiency during data search and exploration.

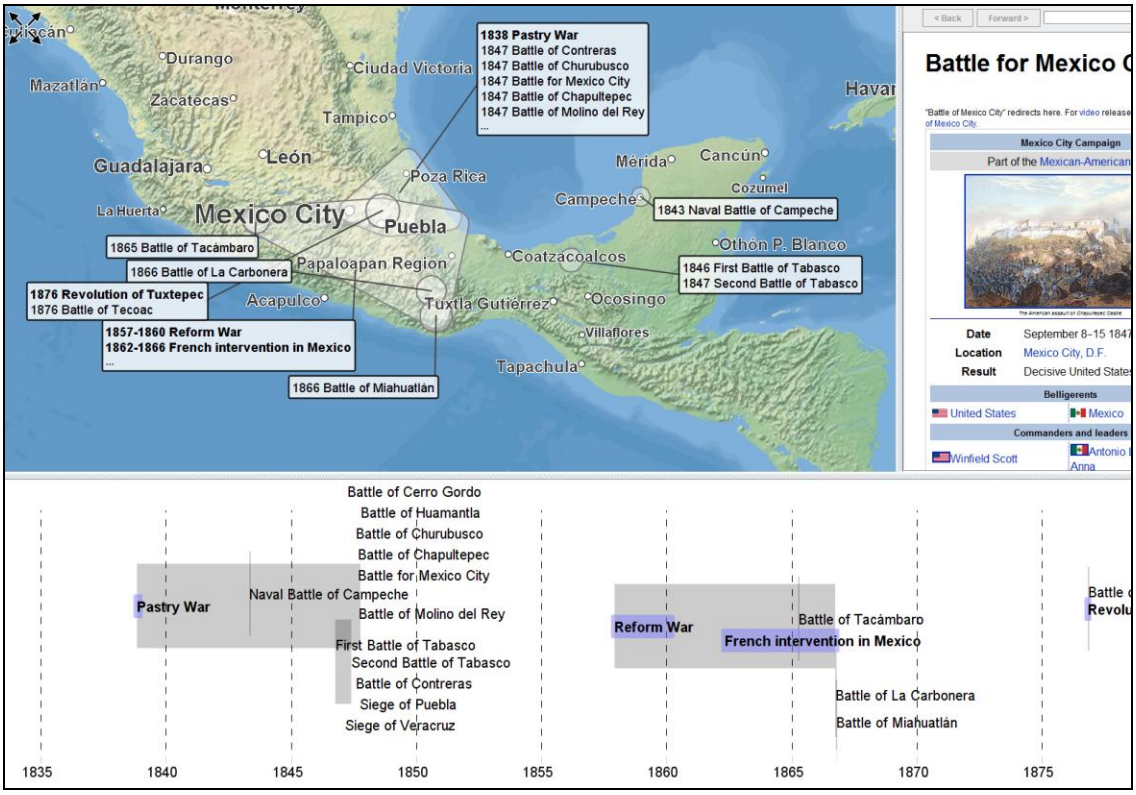

Figure 6. The Mexican History Browser

MaTSE: The Gene Expression Explorer

MaTSE (Craig, Kennedy et al. 2002; Craig and Kennedy 2003; Craig, Kennedy et al. 2005; Craig, Cannon et al. 2010; Craig, Cannon et al. 2012; Craig, Cannon et al. 2013; Kukla 2013), shown in figure 7, combines a variety of visualization and interaction techniques which work together to allow biologists to explore their data and reveal temporal patterns of gene activity. These include a scatter-plot that can be animated to view different temporal intervals of the data, a multiple coordinated view framework to support the cross reference of multiple experimental conditions, a novel method for highlighting overlapping groups in the scatter-plot, and a pattern browser component that can be used with scatter-plot box queries to support cooperative visualization. A final evaluation demonstrated the tools effectiveness in allowing users to find unexpected temporal patterns and the benefits of functionality such as the overlay of gene groupings and the ability to store patterns.

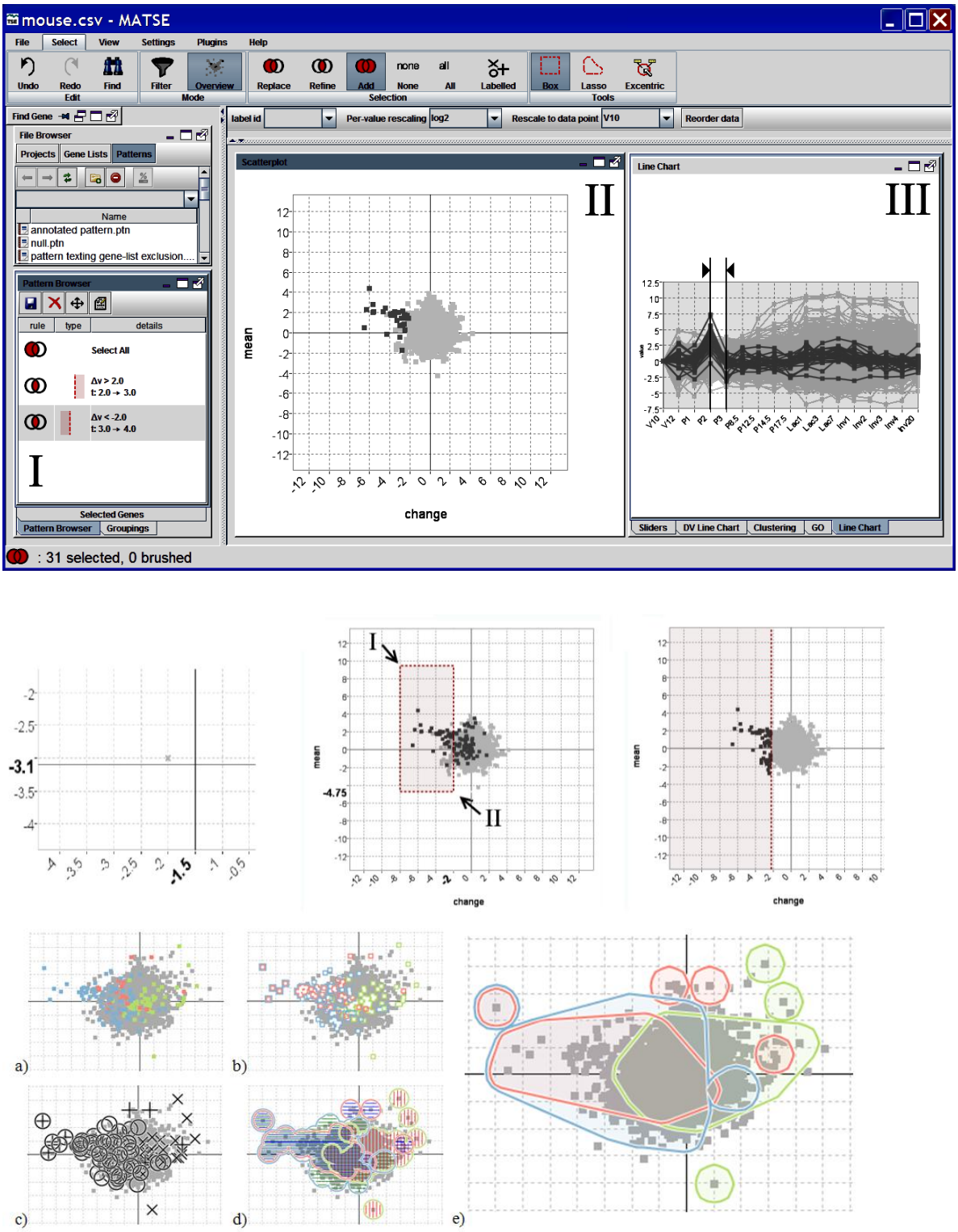

Figure 7. MaTSE: The interface (top), storing patterns (middle) and showing gene groupings (bottom)

Logistics Explorer

The Logistics Explorer application (figure 8) uses two coordinated views of statistical data relating to delivery routes for a small company operating in the Mixteca region of Oaxaca. This allows the company to plan the optimal delivery route and minimize wasted stock. The application runs online within the user's browser and automatically changes with the data source.

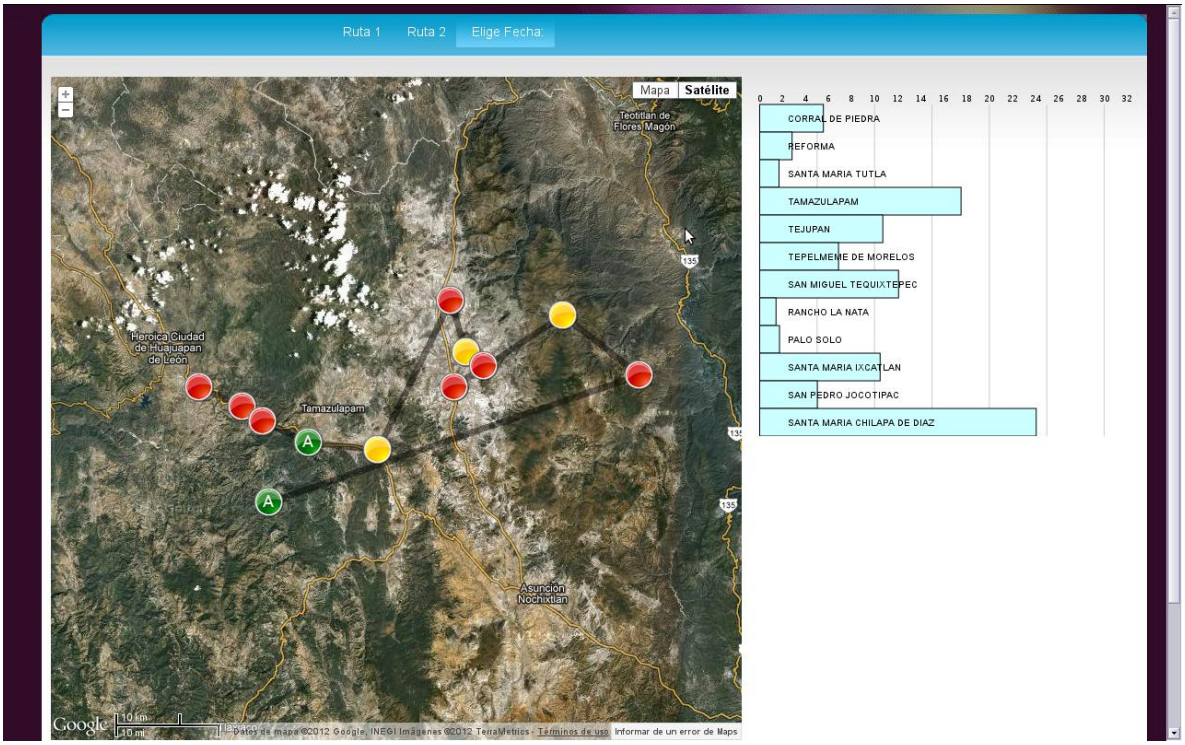

Figure 8. The Logistics Explorer application

Quantile Curve Viewer

Several phenomena are described by two or more correlated characteristics. These dependent characteristics should be considered jointly to be more representative of the multivariate nature phenomenon. Thus, probabilities of occurrence cannot be estimated on the basis of univariate frequency analysis. The quantile function, representing the value of the variable(s) corresponding to a given distribution is one of the most important functions we can study to understand the relationship among different variables. The Quantile Curve Viewer (figure 9) is an application being developed to calculate and show correlations using quantile curves. The advantage of this technique over alternatives such as the use of a

correlation coefficient is that it not only determines the extent of any relationship, but also demonstrates how a dependency is supposed by the existence of data points in opposite quadrants of the graph.

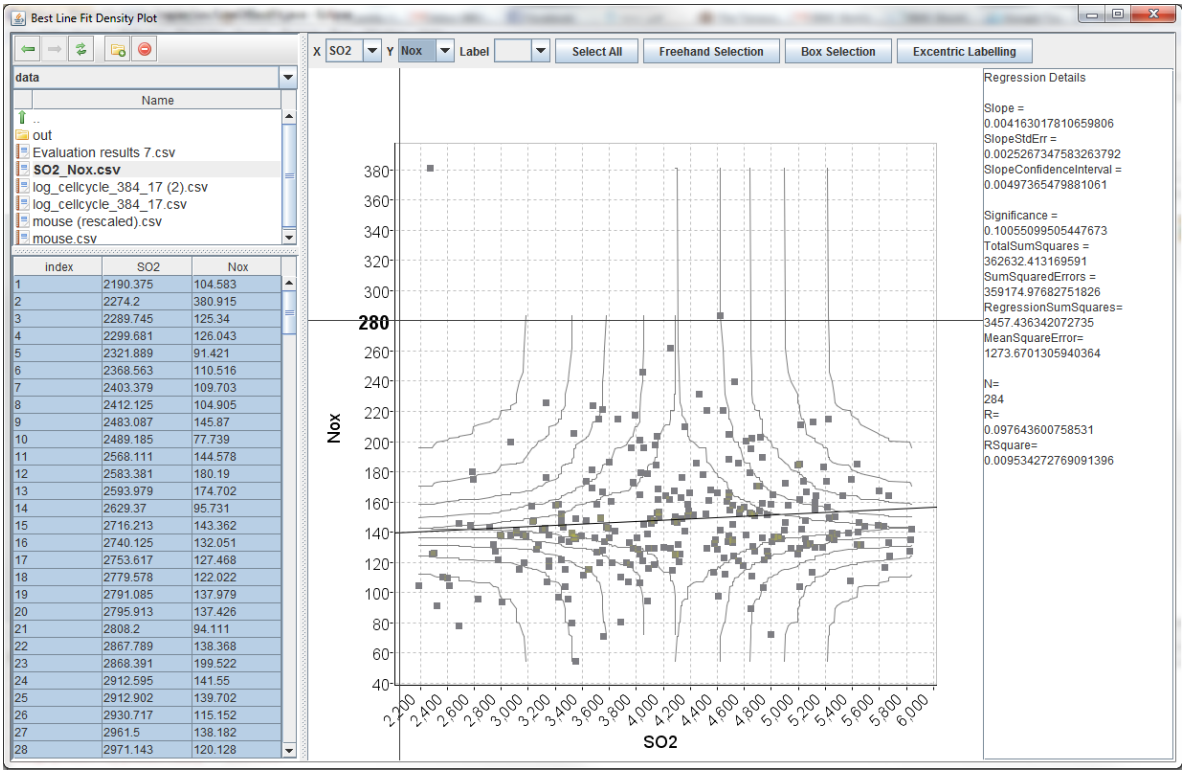

Figure 9. Quantile Curve Viewer

DISCUSSION

Recent advancements in data-acquisition techniques and apparatus have allowed scientists to generate a vast wealth of data with the potential to advance our knowledge in areas such as biology, economics and social sciences. This data tends to be high volume and complex meaning that standard graphical and statistical techniques are often not capable of allowing us to exploit the data to its full potential. It is therefore necessary develop new techniques capable of unlocking the potential of the new data. Information visualization, statistics and mathematical modeling are examples of techniques that can add value to such data. Information visualization is important since it is capable of making high-volume and complex data more accessible to scientists allowing them to develop a better understanding of their data and the phenomena they wish to study. However, information visualization

techniques tend to be highly specialized and can only be developed effectively through close collaboration with usability experts, software developers and scientists.

In fact, the majority of scientific data produced today cannot be properly analyzed without information visualization or at least the application of statistical methods or some form of pre-processing. This means that to be successful in the new age of information intensive science, building interdisciplinary teams and working together is not just important, it's *essential*. Moreover, the quality of the teams we work in, and the quality of the relationships between team members, is vital. We would therefore like to encourage our fellow researchers to look outside their offices, outside their departments, outside their university and outside their state and country to consider the global context of their research not only to find the results and theories that apply to their research but also the people who they can work with to realize the maximum potential of their talents as research professionals. With the growing adoption of internet technologies the world is becoming a smaller place and geographically remote areas such as Oaxaca are no longer remote in terms of global communication. Indeed with its great wealth of untapped talent, previously disadvantaged areas such as Oaxaca are now in the best position to benefit from the new information age by forging links with institutions across the state, country and globe. The adoption of a collaborative interdisciplinary mode of working and an increased adoption of tools for large scale data, such as information visualization, are likely to be key components is this shift toward improving our position in this new age of information intensive science.

CONCLUSION

We have introduced some of the basic concepts of information visualization and demonstrated with our own applications how information visualization can be used to the benefit of scientists in a number of application areas. Information visualization is a great tool to make data more accessible and accelerate the process of scientific discovery. This is particularly relevant in the modern age where the technologies to generate and share large scale data are increasingly prevalent. We hope to continue working towards improving how people are able to work with data by developing more IV techniques and applications. This will necessarily involve establishing academic partnerships and developing existing

relationships with our talented partners in the Sistema de Universidades Estatales de Oaxaca and abroad.

REFERENCIAS

Andrews, K. (1995). Visualising Cyberspace: Information Visualisation in the Harmony Internet Browser. IEEE Symposium on Information Visualization 1995, Atlanta, Georgia, USA, IEEE Computer Society Press.

Bartram, L. (1997). Perceptual and interpretative properties of motion for information visualization. NPIV '97. New Paradigms in Information Visualization and Manipulation, Las Vegas, Nevada, USA, ACM Press.

Bederso, B. B., J. Meyer, et al. (2000). Jazz: an extensible zoomable user interface graphics toolkit in Java. User Interface and Software Technology, ACM Press.

Bertin, J. (1983). Semiology of Graphics: Diagrams, Networks, Maps. Madison, WI, University of Wisconsin Press.

Bryant, B., A. Milosavljevic, et al. (1998). "Gene Expression and Genetic Networks (Session Introduction)." Pacific Symposium on Biocomputing **3**: 3-5.

Card, S. K., J. D. Mackinlay, et al., Eds. (1999). Readings in Information Visualization: Using Vision to Think. The Morgan Kaufmann Series in Interactive Technologies. San Francisco, Morgan Kaufmann.

Card, S. K., G. G. Robertson, et al. (1991). The Information Visualizer, an Information Workspace. ACM CHI '91, New Orleans, Louisiana, USA, ACM Press.

Craig, P. (2012). Visualización de la Información Animada para el Análisis de Sistemas Complejos. Segundo Congreso Mexicano de Ciencias de la Complejidad. Ciudad de México.

Craig, P., A. Cannon, et al. (2010). Pattern browsing and query adjustment for the exploratory analysis and cooperative visualisation of microarray time-course data. Proceedings of the 7th international conference on Cooperative design, visualization, and engineering, Calvia, Mallorca, Spain, Springer-Verlag.

Craig, P., A. Cannon, et al. (2012). MaTSE: The microarray time-series explorer. IEEE Symposium on Biological Data Visualization (BioVis), Seattle, CA.

Craig, P., A. Cannon, et al. (2013). "MaTSE: The gene expression time-series explorer." BMC Bioinformatics (highlights of BioVis 2012).

Craig, P. and J. Kennedy (2003). Coordinated Graph and Scatter-Plot Views for the Visual Exploration of Microarray Time-Series Data. IEEE InfoVis, Seattle, Washington, USA, IEEE Computer Society Press.

Craig, P. and J. Kennedy (2008). Concept Relationship Editor: A visual interface to support the assertion of synonymy relationships between taxonomic classifications. Visualization and Data Analysis 2008, San Jose, CA, Society of Photo-Optical Instrumentation Engineers, Bellingham, WA.

Craig, P., J. Kennedy, et al. (2005). "Animated Interval Scatter-plot Views for the Exploratory Analysis of Large Scale Microarray Time-course Data." Information Visualization **4**(3): 149-163.

Craig, P., J. B. Kennedy, et al. (2002). Towards Visualising Temporal Features in Large Scale Microarray Time-series Data. 6th International Conference on Information Visualisation - IV2002, University of London, London, GB, IEEE Press.

Craig, P. and N. Roa-Seïler (2012). A Vertical Timeline Visualization for the Exploratory Analysis of Dialogue Data. Information Visualisation. Montpellier, France**:** 68 - 73.

Craig, P. and N. Roa-Seïler (2013). A combined multidimensional scaling and hierarchical clustering view for the exploratory analysis of multidimensional data Visualization and Data Analysis. San Fransisco, USA.

Fekete, J. and C. Plaisant (2002). Interactive Information Visualization of a Million Items. IEEE Symposium on Information Visualization, Boston, Massachusetts, USA, IEEE Computer Society.

Holmes, N. (2005). "Explanation Graphics (http://www.nigelholmes.com/)." Retrieved 16 September, 2005.

Kukla, R. C., P. (2013). "Time-series Explorer Webpage (http://www.soc.napier.ac.uk/TSExplorer)." 2005.

Mackinlay, J. (1988). Applying a theory of graphical presentation to the graphic design of user interfaces. Symposium on User Interface Software and Technology, Alberta, Canada.

Mackinlay, J. D., G. G. Robertson, et al. (1991). The perspective wall: detail and context smoothly integrated. ACM CHI '91, New Orleans, Louisiana, USA, ACM Press.

Robertson, G. G., J. D. Mackinlay, et al. (1991). Cone Trees: Animated 3D Visualizations of Hierarchical Information. ACM CHI: Human Factors in Computing Systems, New Orleans, Louisiana, USA, ACM Press.

Seara Vázquez, M. (2009). "Un nuevo modelo de universidad, universidades para el desarrollo." Oaxaca: Universidad Tecnólogica de la Mixteca.

Shneiderman, B. (1996). The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. IEEE Visual Languages '96, Boulder, Colorado, USA, IEEE Computer Society Press.

Stasko, J., M. Guzdial, et al. (1999). Evaluating Space-Filling Visualizations for Hierarchical Structures. IEEE InfoVis, Late Breaking Hot Topics, San Francisco, California, USA, IEEE Computer Society Press.

Tufte, E. R. (1983). The Visual Display of Quantitative Information, Graphics Press.

Tufte, E. R. (1991). Envisioning Information. Cheshire, Connecticut, Graphics Press.

van Wijk, J. J. and W. A. A. Nuij (2003). Smooth and efficient zooming and panning. IEEE InfoVis, Seattle, Washington, USA, IEEE Computer Society Press.

Ware, C., J. Bonner, et al. (1992). "Moving Icons as a Human Interrupt." International Journal of Human-Computer Interaction **4**(4): 341-348.

Ware, C. and G. Franck (1996). "Evaluating stereo and motion cues for visualising information nets in three dimensions." ACM Transactions on Graphics **15**(2): 121-140.

Wertheimer, M. (1961). Experimental studies on the seeing of motion. New York, Philosophical Library.

Yee, K.-P., D. Fisher, et al. (2001). Animated Exploration of Dynamic Graphs with Radial Layout. IEEE Symposium on Information Visualization, San Diego, California.