P. Buono, M. F. Costabile, Insights on the development of visual tools for analysis of pollution data.
In: Proc. of International Conference on Distributed Multimedia Systems (DMS 2012).
Miami, August 9-11, p. 54-59, Skokie, IL 60076, USA:Knowledge Systems Institute, ISBN: 1-891706-32-2

Insights on the development of visual tools for analysis of pollution data

Paolo Buono and Maria Francesca Costabile
Dipartimento di Informatica
Università degli Studi di Bari Aldo Moro
Via Orabona, 4 - 70125 Bari, Italy
{buono, costabile}@di.uniba.it

***Abstract*—Developing visual tools that support data analysis in a specific application domain requires a careful investigation in order to understand needs and expectations of people who will use such tools. The domain experts addressed in this paper are chemists specialized in environmental data analysis. Their main activity is to detect and monitor chemical compounds in the air through many devices in order to detect anomalies or prevent risks. One of the main problems that chemists face is the analysis of the huge amount of data produced by devices. They perform explorative data analysis and are willing to use software tools that can help them to get insights from data. This paper reports the experience in working with chemists to identify interactive visual tools that can be useful for their purposes. It provides insights on the difficulty of creating systems that users find really useful for their work, even when users participate in the design team. Because of the complexity of the considered problem and the fact that people are unable to make explicit all their needs and requirements, the identification of proper tools resulted very challenging.**

Keywords—Environmental data; data analysis; user-centred design; participatory design.

I. INTRODUCTION

Today, one of the problems in data analysis is the quantity of data that have to be analyzed in different application domains and for many purposes. Tools and techniques have been studied and developed in order to assist people in this heavy task. Visual analytics is an emerging interdisciplinary research field that includes, among others, Data Mining and Information Visualization techniques. The purpose is to make sense of very large and complex datasets by combining "automated data analysis with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex datasets" [14]. This research field aims at utilizing the strengths of automatic methods as well as the innate human ability to visually perceive patterns and trends to help people analyzing large complex data.

Many visual analytics tools have already been developed, e.g. Tableau [22], Spotfire [21], Jigsaw [11]. They all claim to be general purposes, i.e. they support users in the analysis process in several context. The experience reported in this paper shows the difficulty to adapt such tools in order to fulfill the requirements of the experts of the application domains in which these tools might be applied, who usually are not IT professionals. Indeed, domain experts are used to their working and reasoning strategies and they are often reluctant to change their habits, and they do not want to be constrained by the technology. The experience reported in this paper shows that, when they are using a technological tool that does not reflect their working habits, after a certain time they do not use it anymore.

We adopted user-centered design and participatory design paradigms in order to design tools that chemists could find adequate to their needs and pleasant to use. User-centered design requires that end users' needs, profiles, tasks, context of use are deeply analyzed and that system is designed and developed by iterating a design-implementation-evaluation cycle, which includes end users in prototype evaluation [10]. Participatory design goes even further in user involvement, since it requires the participation of end users in the design process [23]. The rationale is that users are experts of the work domain thus a system can be effective if these experts are allowed to participate to its design, giving indications on their needs and expectations. However, even by adopting those paradigms, another problem emerged, determined by the fact that end users are unable to make explicit all their needs and expectations during the requirement analysis, even when they are involved in the design team, as suggested by participatory design. As shown in the paper, users really understand how the system works and are able to provide useful feedbacks only when they use the system, or a prototype, in their working practices.

The domain experts addressed in this paper are chemists specialized in environmental data analysis. Everyday, they analyze huge amount of data produced by different devices located on a territory, in order to monitor air pollution. By involving the chemists in a multidisciplinary team, we have analyzed several visual analytics tools that may support the chemists' work. In this paper we report the difficulties encountered in designing tools that satisfy chemists needs and expectations so that they can successfully use them in their working practices.

Next section illustrates the data analysis problem faced by chemists. Section III discusses the attempts made to adapt existing and renowned tools to the chemists' problem, and the reasons they resulted not suitable for them. Section IV illustrates the adopted platform to support the analysis process and the plugins developed to adapt it to the chemists' needs. Section V concludes the paper providing some lesson learned from this experience.

II. THE ANALYSIS OF POLLUTION DATA

Today the air quality is one of the major problems considered by governments. There are different types of pollution that experts analyze. This paper refers to the work of chemists who detect the concentration in the air of specific chemical compounds by analyzing the smell perceived through electronic nose instrumentations, i.e. machines designed to detect and discriminate among complex odors using a sensor array [17]. This array consists of sensors that are treated with a variety of odor-sensitive biological or chemical materials. An odor stimulus generates a characteristic pattern (or smellprint) from the sensor array. Smellprints from known odors are used to construct a database and train a pattern recognition system so that unknown odors can subsequently be classified and identified. Thus, an electronic nose includes hardware components (different devices collecting odors) to collect and transport odors to the sensor array, and electronic circuitry to digitize and store the sensor responses for signal processing. Early prototypes of electronic noses date back to 70's, but only today, thanks to advances in microfabrication techniques, there are detector arrays with enhanced sensitivity and selectivity with characteristics comparable to those of humans.

The chemists addressed in this paper are researchers that are investigating about the causes of unpleasant smells in the air in order to prevent air pollution. They are not people that routinely check pollution data in order to communicate to government about critical situations for human safety, so that proper actions can be decided. Instead, they are searching for reasons of pollution and possible solutions to improve air quality. Their work is not restricted to odor detection, because multiple factors can generate unpleasant smells. In particular, it is important to correlate smell data with some weather parameters coming from weather stations. Weather information is very important to better understand behaviors of the observed chemical compounds because they are detected in the air. The most relevant weather parameters are wind speed and direction.

Although different companies are now producing electronic noses these systems are not yet totally reliable. Researchers want to check the responses of the electronic system with those provided by human beings. To this aim, they have selected sixteen housewives, resident in different locations of the territory, who are asked to note in a form every time they feel a bad smell in the air, annotating the time, the duration and the intensity of the smell. The ladies transmit such data via phone to a toll free number. Figure 1 shows the different ways of collecting data about air pollution, due, for example, to the emissions of a landfill. Besides the already mentioned data collected by the electronic nose, the housewives and detectors of weather parameters, data are also collected by PID (PhotoIonization Detector) and OPC (Optical Particle Counter), two more devices to collect information about pollution.

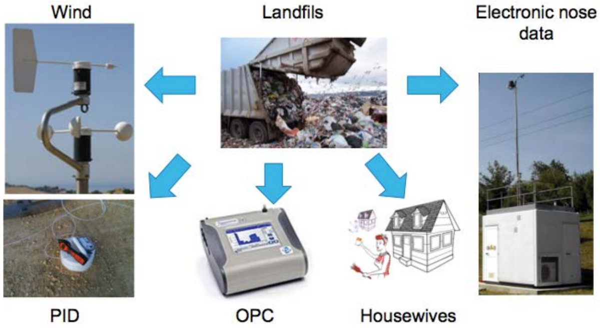

Figure 1. Sources of analyzed data

Different devices of the electronic nose, located in different stations on the territory, monitor chemical compounds in the air with different techniques, time granularity, and unit measures. Data are stored in logs files, which can be transferred via Internet. Chemists collect such log files, integrate and/or compare them with data coming from other sources (weather, housewives), apply transformations and analyze the resulting data. Before the beginning of this work, for their analysis, they used tools, such as MATLAB and Microsoft Excel, in order to make computations, and/or produce visualizations that may help them to understand and to explain what is going on, and to perform predictions.

III. PROBLEMS IN USING VISUALIZATION TOOLS

There are many visualization tools in literature with different purposes. Several authors have also provided some classifications. For example, Shneiderman has classified visualizations tools according to tasks that users have to accomplish and according to data types [20]. From the description provided in the previous section, it is evident that the data considered in this paper are time dependent. All considered data (coming from the electronic nose devices, weather parameters, data reported by the selected housewives) are associated to the time when they have been collected. Thus, they fall in the category of time series. In fact, a time series is a sequence of data points, measured typically at successive time instants spaced at uniform time intervals. Sometimes the constraint of uniformity in collecting time series data is relaxed, but not in the domain presented in this work, so this is not addressed in this paper.

There is a wide literature in time series and in time dependent analysis. A survey is provided by Aigner et al. [1], who analyze a wide spectrum of key techniques in visualizing time oriented data, classifying each of the about one hundred identified techniques according to different dimensions; one refers to the main tasks users want to perform for analyzing time series data, namely: select, explore, reconfigure, encode, abstract/elaborate, filter, connect, undo/redo, and configuration change. These tasks match very closely the needs observed during the chemists' requirements analysis.

We setup a multidisciplinary team, which included the domain experts, in order to analyze different tools, looking for those that might better fit the experts' needs. The team considered various tools. One that was deeply discussed is LiveRAC [15]. One of it's main characteristics is that it displays on the screen many time series using a focus + context technique combined with semantic zoom [18]. Both are useful techniques that allow focusing on specific time series, keeping the context in which they are visualized, which is represented with much less details. However, chemists were not enthusiast about this kind of interface, since they found the screen too cluttered and, in their opinion, too many details were hidden by the visualization. Moreover, they found the side-by-side comparison of time series not useful for them because they are used to compare overlapping graphs.

Based on their comments, TimeSearcher was analyzed [5], [6]. TimeSearcher provides users with interactive visualization of multivariate time series in a single screen. For example, the screen in Figure 2 shows 8 panels, each displaying time series of a variable (e.g. temperature) whose values are collected at different locations. Panels are vertically juxtaposed for an easy comparison among variables. Data are aligned according to the time, and a vertical blue line (on the left in each panel of Figure 2) indicates this alignment. TimeSearcher supports explorative analysis of time series, to this aim it allows user to search for interesting patterns in the series and to filter according to some specific data values and time intervals. A detailed demo was shown to the domain experts, during which the features of TimeSearcher were discussed. The experts appreciated the tool and, in particular, the possibility to compare different time series in the same panel in order to easily identify correlations. For example, as shown in Figure 2, the interface of TimeSearcher has 8 panels; each panel visualizes several time series of the same variable, e.g. temperature, recorded at several monitoring stations.

The chemists decided to adopt this tool, since they were convinced that it was much better than spreadsheets in permitting visual interaction with data, thus supporting a more effective analysis. In order to manage the different formats of the data collected from the heterogeneous sources, a software module has been developed to convert the data in the format required by TimeSearcher. This module requires a human's intervention to solve some ambiguities in the data. This process required time, but this was not considered a problem since data conversion did not occur so frequently.

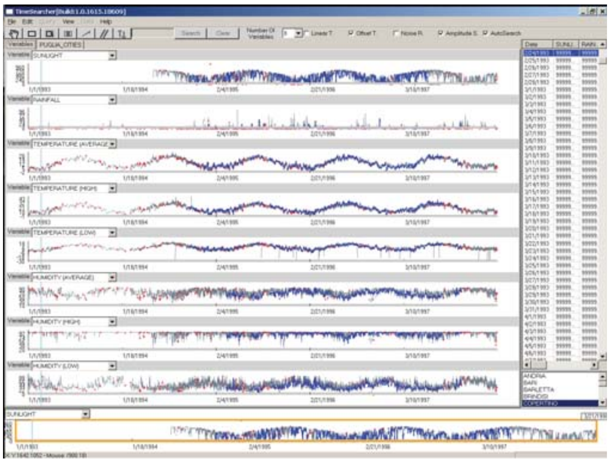

Figure 2. A screenshot of TimeSearcher with 8 panels, each displaying time series of a variable, whose values are collected at different locations.

Chemists used TimeSearcher, appreciating most of its features. However, pretty soon some of them complained about TimeSearcher, because they primarily considered a different type of analysis, i.e. they concentrated on data of a single monitoring station with more than 12 devices, each producing data about a specific variable. Such chemists wanted to plot values of these variables in the same panel. TimeSearcher does not permit this. It is worth noticing that this requirement never became explicit during the discussions with chemists and their trials with demos of TimeSearcher usage. It actually emerged only once they worked with the system during their daily activities. This finding has been already highlighted in [24] and experienced also in [3]: "end users provide the most valuable feedback about the possible problems once they get to work with the new system in real settings".

Another drawback emerged, this time due to a new requirement generated from the evolution of the chemists' activities. In fact, these chemists are experts who carry out research work related to the air quality. The data they work with are not collected routinely, as done in centers devoted to the air pollution control. However, at a certain point they started to collect data much more frequently than before, and this implied to perform much more often the data conversion process to adapt data to TimeSearcher. As we said, this process needs the involvement of a researcher. Thus, chemists very soon got annoyed of repeating the data conversion of new data so many times. As a consequence, they stopped using TimeSearcher and went back to their original work with MS Excel and MATLAB that, they also said, allowed them to perform many other computations required from their recent research directions.

These new requirements shifted the interest of the participatory team to the data pre-processing phase, also because they started collecting data from other devices, namely PID and OPC (see Figure 1), which were not considered before. As it will be shown in the next session, data preprocessing is one of the main phases of the overall data analysis process because, especially when data come from different sources and are in different formats, as in our case, data have to be manipulated to get them in a form usable by the visualization tools, and check their quality.

IV. THE ANALYSIS PROCESS WITH KNIME

Our activity focused on the preprocessing phase due to the new requirements of the chemists, who often had to analyze new data. Preprocessing includes all the necessary steps to prepare data for the analysis. Some authors now refer to these steps as data wrangling, which is considered one of the big issues for a meaningful data analysis. Data wrangling is defined as "a process of iterative data exploration and transformation that enables analysis" [12]. In fact, an analyst has to first diagnose the data to make sure that they are usable for the analysis questions, to see the efforts required to put them into an appropriate format for the analysis tools, to check their quality, i.e. if there are missing data, inconsistent values, unresolved duplicates, etc. In other words the data have to be transformed and cleaned into a usable state. Data wrangling is actually the process to make data useful.

Numerous techniques for cleaning and integrating data have been proposed within the database community (for example see [9],[13],[19]). However, many of them focus on specific data quality problems, are not interactive, or are not accessible to a general audience. One of the current challenges in this field is to enrich data processing technologies with innovative visual interfaces that support data diagnostics and transformations [12]. Going back to the pollution data analysis, we realized that we needed tools for a more accurate data preprocessing and that we had to shift our attention to the overall analysis process.

A. *The Knime platform*

We looked at platforms that support this analysis process, since several are now available. However, most of them concentrate on some phases of the process. We found that Knime (Konstanz Information Miner) provides a working environment that supports users in their activities by providing several tools useful for the phases of data preprocessing, analysis, and exploration [7]. Developed at the University of Konstanz, it is open-source and based on Eclipse.

In Knime, the user creates the workflow for his/her overall data analysis process. Figure 3 shows an example. The user has composed the workflow by selecting, from a menu showing all available tools, those necessary for his/her analysis. Each node represents a step of the process; in other words, a node is a plugin that performs a specific activity. As shown in Figure 3, the node File Reader, on the left of the figure, loads data in the environment. Then data are transferred to the K-Means node, which creates clusters of data. Note that the little blue square on the right of a node means that the node is able to display data. The white arrows represent communication ports; if the arrow points toward the node it is an input port, otherwise it is an output port. K-Means node is connected to a Color Manager node, which associate colors to the clusters created by the previous node. The Color Manager node is connected to two other nodes that visualize data using different techniques, according to colors set in Color Manager. In Figure 3 the user is visualizing the menu associated to the Scatter Plot node and is about to press the "View: Scatterplot" menu item. In the example of Figure 3 all nodes are marked with a green light (at the bottom of the node icon) because the analysis is finished without errors. Each node can be in one of the following three states: to be configured (red light), ready to run (yellow light), node processed (green light).

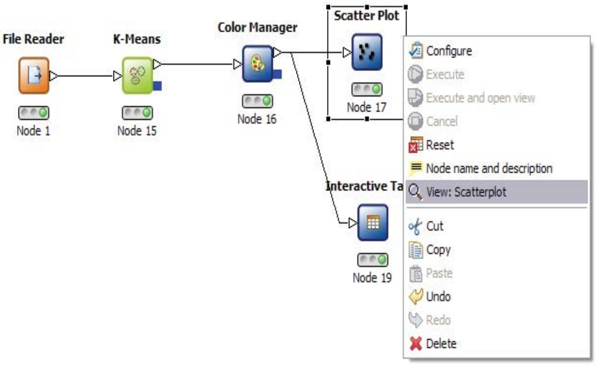

Figure 3. A Knime workflow with five nodes. File reader loads a data file and sends data to the K-Means clustering algorithm. The Color Manager node associates colors to clusters and the output is visualized through both a scatter plot and an interactive table.

Working with Knime, our participatory team soon realized that, even if Knime includes many tools, it was necessary to develop other plugins that could process the specific data of the chemists' problem. In order to create as soon as possible a running prototype with which chemists could work, we developed three plugins. These are:

1) Multifile reader, which loads many files containing data produced by a single data source (like those represented in Figure 1) and merges them according to the time of sampling;

2) Nose-wind converter, which merges electronic nose data with weather parameters;

3) PID/OPC converter, which merges PID and OPC data with weather parameters.

B. *Preliminary formative test*

A user test was performed as part of our formative evaluation. This refers to evaluation activities that are carried out during the design and development cycle of a software product in order to get feedback that it is very useful in "forming" the product, so that it can really fit users' needs and requirements. Sometimes only inspection methods are used in formative evaluation, since tests with users are more resource demanding. However, our experience is a further indicator that formative evaluation requires that users test the prototypes, especially in cases, like ours, in which the system is complex and its target users are professional people, who are very demanding in terms of functionality the system provides and expect a good mapping of the interaction strategies available in the system with their habits and ways of solving their problems.

The user test was conducted using the thinking aloud method, which is well known and appreciated for providing very useful results without requiring many resources [16]. Four domain experts participated. They had experience in the use of their traditional analysis tools, i.e. MS Excel and MATLAB. They were asked to perform some preprocessing activities with Knime so that they had to use the developed plugins. Specifically, the tasks covered activities ranging from the data acquisition to the data visualization.

The test took place in the participants' office. The four participants were first together in a session during which one of the computer scientists did a short presentation of the platform (see Figure 4). Then each participant performed the test in her/his office, sitting at a desk with two evaluators: the facilitator and the observer. A laptop was used for the test, equipped with a screen recorder, which recorded the interaction performed by the user, and a webcam to observe the user's reactions. In accordance with the thinking aloud method, the participant was asked to "think aloud", in order to get the most information on the strategies employed by the participant to accomplish the tasks and better understand any possible difficulty.

Figure 4. Training phase of the test

In Figure 5 a test session is shown. The role of the facilitator, sitting next to the participant, is to assist her in the event of difficulties and encourage to move forward in testing, but without interfering with the task execution, and without providing or requesting information that can influence participant's behavior. The observer was sitting next to the facilitator and his role was to observe the test, also annotating in a table the time taken for each task. After the test, each participant was interviewed in order to get qualitative feedback.

Figure 5. A participant performing a task

To the purpose of this paper, it is useless to report the details of the test results. It is enough to say that the chemists easily understood how to work with Knime and appreciated its utility for their work, especially because it allowed them to better understand the overall process and to easily repeat the preprocessing activities. They also complained about some usability problems of Knime that, however, will not compromise the use of the platform because they can be solved by providing some initial training. In this case, this is not so bad since the chemists will use the system very frequently and they will become soon expert users.

V. Conclusions

This paper presents our experience of working with chemists in a multidisciplinary team for developing visual tools that support the analysis of pollution data. The experience indicated some problems that have to be solved for creating visual analytics tools that are useful and usable for their intended users. A first problem is that it is very difficult to adapt visual tools that are claimed to be general in contexts that take into account very complex and very specific problems, like the one addressed in our case study. Another problem is that data preprocessing is extremely important in visualization systems and need special care [12]. The two problems refer to the development of any interactive system, and not only to visual analytics systems: 1) it is extremely hard for the users to communicate their requirements, even if a participatory design approach is adopted; 2) the only way to evaluate if a software system is useful and usable for its users is to let them to put their hand on prototypes.

Our suggestion to overcome these problems is to perform evaluations with methods involving end users, observing them when they are working with running prototypes of the system, as the user test described in this paper. We plan to go even further, by requiring longer sessions of use of the prototype by testers while working in their daily activities. This is in line with other findings already provided in literature [3],[24].

Future work could consider the design of user interfaces that support the analysis process on mobile devices. Indeed, it could be useful to perform some analysis steps at the locations where the devices are installed. This poses a great challenge in data visualization due to the limited size of the screen, which requires the adoption of specific visualization techniques, such as overview + details and semantic zoom (see, for example, [2], [4], [8]).

REFERENCES

[1] W. Aigner, S. Miksch, H. Schumann, and C. Tominski. Visualization of Time-Oriented Data. Springer-Verlag New York, 2011.

[2] C. Ardito, P. Buono, Costabile M., R. Lanzilotti (2006). Two Different Interfaces to Visualize Patient Histories on a PDA. Proc. of MobileHCI 2006, Espoo, Finland. September 12-15, 2006. New York: ACM Press (United States), pp. 37–40.

[3] C. Ardito, P. Buono, M. F. Costabile, R. Lanzilotti, A. Piccinno, A. L. Simeone. Analysis of the UCD process of a web-based system. Proc. of DMS 2010. Oak Brook, Illinois, USA, 14-16 October, 2010, Skokie, Illinois: Knowledge Systems Institute, pp. 180–185.

[4] B. B. Bederson, A. Clamage, M. P. Czerwinski, and G. G. Robertson. DateLens: A fisheye calendar interface for PDAs. ACM Trans. Computer-Human Interaction, vol 11, no. 1, 2004, pp. 90–119.

[5] P. Buono, A. Aris, C. Plaisant, A. Khella, B. Shneiderman. Interactive Pattern Search in Time Series. Visualization and Data Analysis. Proc. of VDA 2005, 16-20 January 2005, San Jose, CA USA, SPIE, Washington DC, pp. 175–186.

[6] P. Buono, C. Plaisant, A. Simeone, A. Aris, G. Shmueli, and W. Jank. Similarity-based forecasting with simultaneous previews: A river plot interface for time series forecasting. Proc. of IV 2007, Washington, DC, USA, 2007. IEEE Computer Society, pp. 191–196.

[7] N. Cebron and M. R. Berthold. Adaptive active classification of cell assay images. Knowledge Discovery in Databases: PKDD 2006, volume 4213, Springer Berlin / Heidelberg, 2006, pp. 79–90.

[8] L. Chittaro. Visualizing Information on Mobile Devices. Computer vol. 39, no. 3, 2006, pp. 40–45.

[9] A.K. Elmagarmid, P.G. Ipeirotis, and V.S. Verykios. Duplicate record detection: A survey. IEEE Trans Knowl Data Eng 2007; 19(1), pp. 1–16.

[10] International Organization for Standardization, ISO 13407: Human-Centered Design Process for Interactive Systems, 1999.

[11] Jigsaw. http://www.cc.gatech.edu/gvu/ii/jigsaw/. Last accessed: April 2012.

[12] S. Kandel, J. Heer, C. Plaisant, J. Kennedy, F. vanHam, N. H. Riche, C. Weaver, B. Lee, D. Brodbeck, and P. Buono. Research directions in data wrangling: Visualizations and transformations for usable and credible data. Information Visualization Journal, vol. 10, no. 4. 2011. SAGE, pp. 271–288.

[13] H. Kang, L. Getoor, B. Shneiderman, M. Bilgic, and L. Licamele. Interactive entity resolution in relational data: A visual analytic tool and its evaluation. IEEE Trans Visual Comput Graph 2008; 14, pp. 999–1014.

[14] D.A. Keim, J. Kohlhammer, G. Ellis, F. Mansmann. (Eds), Mastering the Information Age - Solving Problems with Visual Analytics. Eurographics Association. 2005.

[15] P. McLachlan, T. Munzner, E. Koutsofios, and S. North. Liverac: interactive visual exploration of system management time-series data. Proc. of CHI 2008, New York, NY, USA, 2008. ACM, pp. 1483–1492.

[16] J. Nielsen. Usability Engineering, Academic Press, San Diego, 1993.

[17] T.C. Pearce, S.S. Schiffman, H.T. Nagle, J. W. Gardner. (Eds) Handbook of Machine Olfaction: Electronic Nose Technology. Weinheim: Wiley-VCH Verlag. 2003.

[18] R. Rao, and S. K.Card, The table lens: merging graphical and symbolic representations in an interactive focus+context visualization for tabular information. Proc. of CHI '94. ACM, New York, NY, USA. 1994, p. 222.

[19] G.G. Robertson, M. P. Czerwinski, and J. E. Churchill. Visualization of mappings between schemas. Proc. of CHI 2005, pp. 431–439.

[20] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. Proc. of Visual Languages (Boulder, CO, Sept. 3–6). IEEE Computer Science Press, Los Alamitos, CA, 1996, pp. 336–343.

[21] Spotfire. http://spotfire.tibco.com/. Last accessed: April 2012.

[22] Tableau. http://www.tableausoftware.com/. Last accessed: April 2012.

[23] D. Schuler, A. Namioka, Participatory Design: Principles and Practices, Lawrence Erlbaum Associates, Inc., 1993.

[24] J.L.Wagner, G. Piccoli, Moving beyond user participation to achieve successful IS design. Communications of the ACM, vol. 50, no. 12, 2007, pp. 51–55.