Evaluating Visual Table Data Understanding

Nathalie Henry
INRIA Futurs/LRI/University of Sydney
Bât 490, Univ. Paris-Sud, 91405 Orsay, France
Nathalie.Henry@lri.fr

Jean-Daniel Fekete
INRIA Futurs/LRI
Bât 490, Univ. Paris-Sud, 91405 Orsay, France
Jean-Daniel.Fekete@inria.fr

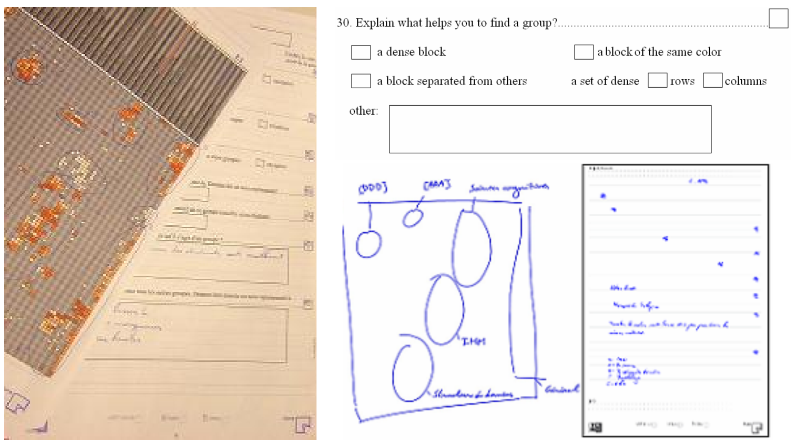

Figure 1: Evaluation using Anoto pen and Digital paper.

ABSTRACT
In this paper, we focus on evaluating how information visualization supports exploration for visual table data. We present a controlled experiment designed to evaluate how the layout of table data affects the user understanding and his exploration process. This experiment raised interesting problems from the design phase to the data analysis. We present our task taxonomy, the experiment procedure and give clues about data collection and analysis. We conclude with lessons learnt from this experiment and discuss the format of future evaluation.

Author Keywords
Information Visualization, Visual Table Data, Evaluation, Controlled Experiment

ACM Classification Keywords
H.5.2 [User Interfaces] Graphical User Interfaces – Evaluation/Methodology

INTRODUCTION
Table data are widely used to collect and analyze numerical values. Spreadsheet calculators popularized the use of table data and offered numerical analysis tools, information visualization systems such as Infozoom[7] or TableLens[5] developed visual and interactive interfaces to explore table data. However evaluating how these systems support the user exploration and quantifying how they improve his understanding still remains a challenge.

Jacques Bertin was the first to introduce visual table data and to attempt to formalize what is table data understanding [2]. He showed that the key to understand visual table data is a matter of ordering its rows and columns, i.e. its layout. To evaluate the layout performance, he defined understanding as the ability to answer three crucial questions about data: what is the data? what are the groups? what are the outliers? Number of related works extend these tasks; [8] presents a classic taxonomy of cognitive tasks, [6] proposes a task-by-data-type taxonomy and more recent work defined an analytic task taxonomy [1]. We selected important tasks for visual table exploration and structured them into a three-level hierarchy. We used this task hierarchy to conduct a controlled experiment aiming at evaluating how visual table data layout affects user understanding. We use questionnaires printed on digital paper and anoto pen to record subjects' answers (Figure 1). In this paper, we describe the design of this evaluation and explain the problems we encounter.

CONTROLLED EXPERIMENT
Our goal is to measure if visual table data layout affects the user understanding and how it affects it. This experiment is a first step toward the evaluation of interactive information visualization systems. We

primarily focused on static visualizations and avoided all interaction issues. We designed this experiment as a 3 layouts*2 datasets within subject. We preferred a within-subject design to limit the inter-subject variability, which is the difference between the exploration processes of two different subjects. To conduct this experiment, we used three table ordering algorithms commonly used to reorder rows and columns of table data. The first order is alphabetic (A) and considered as a control order. The two others (C) and (T) are issued from the field of bioinformatics. (C) is a hierarchical clustering followed by a linearization and is presented in [3]. (T) is based on the traveling salesman problem and described in [4].They are used to reorder DNA microarrays (visual table data presenting gene expressions). We chose them based on their robustness and scalability: they can handle real datasets with noise up to ten thousands of elements in a few seconds.

DATASETS

We chose two real datasets in order to conduct this experiment in a realistic context. Our goal was to find interesting datasets of reasonable size to keep our subjects motivated. Motivation is a crucial factor as we intend to evaluate how a user explores and understands a dataset. This experiment had to keep subjects interested in the data and willing to explore its structure to get realistic results.

As our subjects were mainly computer science students and researchers, we chose the university master's grades dataset (Figure 2) in which they appeared as students or teachers. We picked also a more general table data issued from the CIA World Factbook (Figure 3). This dataset gives statistics (productivity, export, population) for most of the countries of the world.

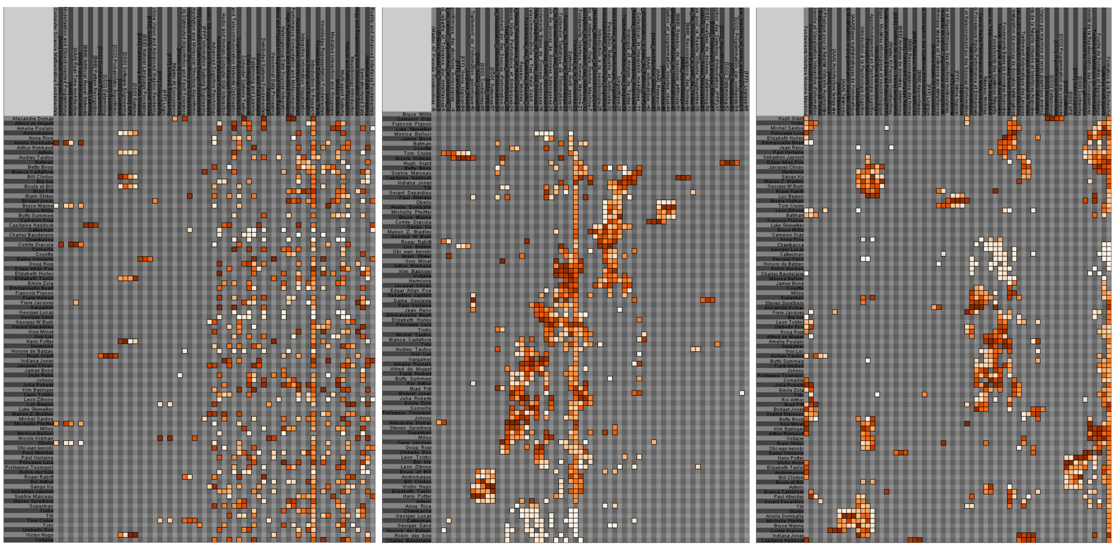

Figure 2: Visual table data of the university master's grades: 88 students(rows)*65 courses(columns). The color intensity represents values (low values are white, high values are red, grey represents no value). Alphabetic order (left), automatic orders C (middle) and T (right).

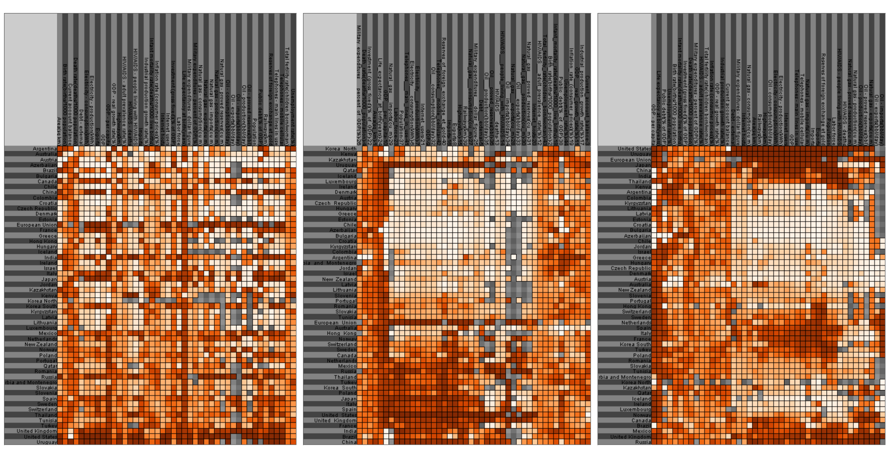

Figure 3: Visual Table Data for World Factbook 2005 (filtered): 55 countries(rows)*44 criteria(columns). The color intensity represents values (low values are white, high values are red, grey represents no value. Alphabetic order (left), C order (middle) T order (right).

TASKS

Firstly, we considered using interviews to collect user findings and comments on the fly. However data analysis would have been perilous especially because of the inter-subject variability and therefore results can be arguable. Instead, we proposed to express the exploration process in terms of tasks, defining what lies under "understanding table data". After several interviews with novice users and visualization experts, we ended up with a set of tasks and we organized them hierarchically.

We organized this set in three complexity levels. We also distinguished readability from interpretation (Figure 4). Readability tasks deal with the visual representation and are independent of the dataset. In other words, subjects can perform readability tasks without interpreting the representation. Understanding a visual table is a combination of readability and interpretation.

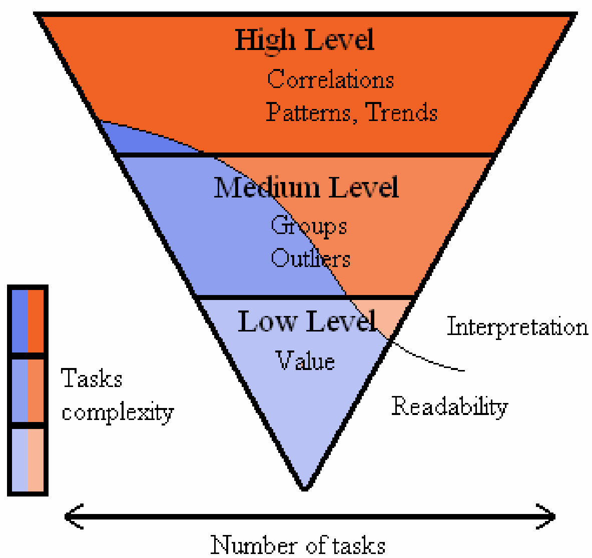

Figure 4: Understanding visual table data, a hierarchy of tasks.

DATA COLLECTION

Subjects had to answer a 4 page questionnaire using an Anoto digital pen on Anoto paper. With this technology, users can read and draw on regular paper with an almost regular pen. The strokes drawn on paper are captured by the pen and can be sent to a computer. In addition to the stroke trajectory, Anoto pen also identify pages and provide the precise time of each stroke. We chose Anoto technology to simplify the tasks and perform the evaluation without using a computer, in a comfortable setting for reading matrices.

To limit the experiment duration, we choose a subset of tasks. Table 1 presents master's grades tasks. When we design the questionnaire, we kept in mind two goals: do not orient the exploration and provide tasks from basic to complex.

Subjects had to tick boxes for multiple choice questions, write comments and explanations for open questions and draw circles or annotations directly on the visual table printed on a separated page. The digital pen recorded the time of each stroke and the stroke location on the digital paper. Thus, we collected quantitative and qualitative data.

Readability	**Interpretation**
Low Level: on representation visual coding	
Find a specified student or course	Find the course where students performs better/worse
Give the grade for a given student and course	Find the student with the highest/lowest grades
Find how many students followed a given course	
Medium Level: on groups and outliers	
Circle a group	Give it a meaningful name
Give a representative element of the group	Give a representative element of the group
Circle outliers	Explain why they are outliers
High Level: on correlations and trends	
Give two courses correlated/ uncorrelated	Give two courses correlated/ uncorrelated
	Give the main trends of this master (in terms of topics)
	Explain how students/courses have been ordered

Table 1: Tasks for the master's grade dataset.

DATA ANALYSIS

The questionnaire contained 32 questions and a visual table representation. Half of the questions were multiple choice questions, the other half were open questions. For each, we collected the time of answer. We analyzed it using a classic analysis of variance. We also collected error rates for quantitative data (multiple choices questions). We develop a strategy to define correctness for more qualitative data such as open questions, comments or explanations: add multiple choice questions after. In addition to time and errors, we collected labels and drawings directly written on the visual representation. The related tasks were: circle groups, give them a meaningful label, and circle outliers. To analyze this data, we used visualization (Figure 5). Preliminary results are very interesting.

We can observe a consensus among all subjects especially for the C layout but also for several groups of the T layout. In addition to this visual data, we analyzed labels and classified them into categories in order to determine first if groups found by most of the subjects have the same connotation, then to compare groups from one layout to another. Although some of the results are not analyzable, our primary analysis points out notable facts. For example, most of the subjects identified a group of students attending cognitive science courses. This group is circled in Figure 5. They labeled it "cognitive sciences", "cognition", "cognitive psycho.", "psychology" or even "social sciences". It represents a real trend in the data as this master course is the result of a fusion of three previous masters, one being the cognitive sciences master. The interesting observation is that subjects identified this group in both C and T layouts although it was split into three in the T layout. That means that the concept of group is not only a *visual pattern* but also a matter of *interpretation*. This explains the importance to distinguish readability from interpretation while studying the understanding process.

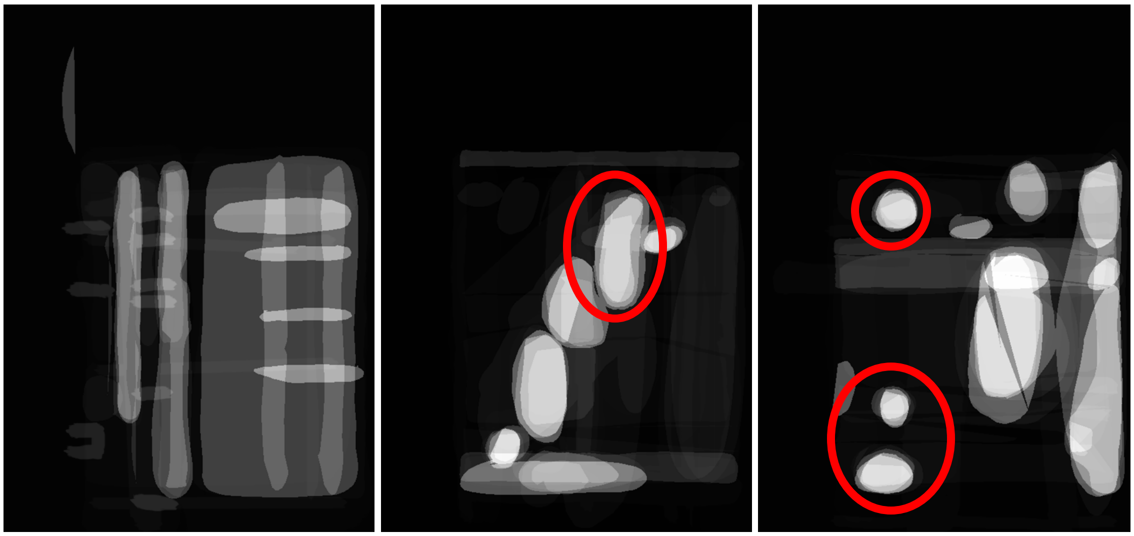

Figure 5 : Superposition of groups for all subjects. Master's dataset. Alphabetic order (left), C order (middle) and T order (right). A group of cognitive science students (identified by our subjects) is circled in red.

LESSON LEARNT: KEEP MAXIMUM CONTROL

Keeping control in a realistic setting is always a key issue for experiments. In this section, we present few lessons we learnt during this experiment.

Don't use real datasets

We encountered problems to create relevant questionnaires for real datasets. As we did not generate the datasets, our main concern was to find the "truth" about the data, *i.e.* determine what its structure was. We decided to collect findings from a set of users (novice and experts) and consider the union of all findings as the "truth". Concerning the master students' grades, we organized a meeting with teachers and administrative staff. Then, we asked questions to several students. For more general dataset, we chose several users both computer science researchers and students.

However, we argue against the use of real datasets, essentially because what we consider being the structure of the data is highly arguable. A safer approach (for analysis) would be either to modify real datasets and increase trends or pattern in their structures or to generate datasets. Then, we would be able to *control the results.* On the other hand, it is no more a realistic context.

Don't ask too much

When dealing with exploration or understanding evaluation, it is important to reduce the duration of the experiment to keep subjects involved and willing to explore. Reducing the number of different tasks but having them performed several time may also help reduce the variability. This is particularly important for complex tasks.

Don't exclude free comments or open questions

Analyzing free comments, explanations or open questions is a nightmare. Our solution is to back open questions with multiple choices questions. After running a few pilot experiments and collecting various answers to open questions or comments, we categorized them and elaborated multiple choice questions. We first ask the open questions and later in the questionnaire the multiple choices questions. Thus, analysis is much simpler and we still have an opportunity to collect original answers.

Don't leave too much freedom

We used digital pen and paper to avoid interface artifacts on user's understanding. However, the resulting artifact on the experiment was maybe worse as we left users the possibility to answer questions out of order. Therefore, we argue for more control using limited software instead of pen and paper.

Don’t underestimate training

During this experiment, we observed two interesting behaviors: the “answer-even-if-you-don’t-know-behavior” and the “finish-as-fast-as-possible-behavior”. These behaviors are probably questionnaires artifacts.

However, training should probably helps with the “answer-even-if-you-don’t-know-behavior”. All tasks provided during the training had an almost obvious answer. Providing the user a question impossible to answer and explain him that he did not have to answer could have help to prevent this behavior. Therefore, do not underestimate training but also pilot experiments which help you to detect unusual behaviors.

Concerning the “finish-as-fast-as-possible-behavior”, we will limit the time for each task and reduce the duration of the experiment.

CONCLUSION

In this paper, we presented an experiment attempting to evaluate how visual table layout affects user understanding. We presented the design of our experiment, gave clues about data analysis and present the lesson we learnt: *keep maximum control*.

This work is in progress and we still are studying how to analyze some of our data. For example, during the experiment debriefing we asked our subjects to sketch what they recalled of the visual tables (Figure 6). We have no clue how to effectively get significant results from this data.

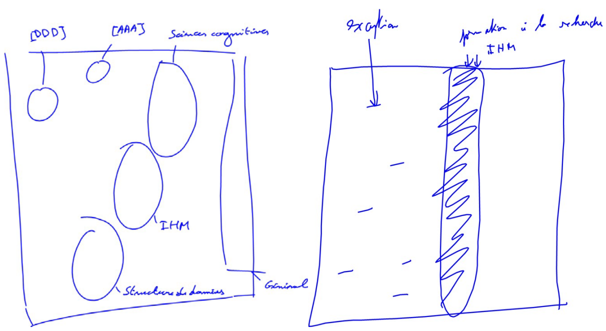

Figure 6: Memorization sketches for master’s grades. We may roughly notice than the left drawing is a combination of C and T layouts whereas the right drawing seems to merge A and C orders. How could we deeper analyze these sketches?

Before concluding, we would like to discuss the controlled experiment design dilemma and raise a deeper question about the evaluation format. Designing an experiment as within-subject or between-subject is a classic dilemma. In this experiment, this means choosing between validity of collected data (which decreases significantly when the subject is not motivated) and inter-subject variability (variability between the exploration processes of two individuals). In our case, within-subject design was really painful for the subjects because of the length of the experiment (2*1h30) and validity of data collected at the end of the experiment is therefore arguable. Would between-subject design have been preferable? We observe strong differences between subjects’ exploration processes especially in matter of answer times. Are those differences only due to subjects motivation? How to limit the inter-subject variability?

Finally, we noticed the problem of controlling without leading which leads to a deeper question: are controlled experiments the right format to evaluate representation understanding? A longitudinal study would solve the motivation problem as subjects would use their own datasets. However it raises other questions: how to collect data? How to distinguish readability from interpretation? How to validate or compare results?

To conclude, this controlled experiment is only the first step towards the evaluation of information visualization systems dedicated to exploration and analysis. In this experiment, we focused on static visualizations for visual table data. Our future work will include interaction and explore different evaluation formats such as longitudinal studies.

REFERENCES

1. Amar, R., Eagan, J. and Stasko, J., Low-Level Components of Analytic Activity in Information Visualization. in *IEEE Symposium on Information Visualization* (Minneapolis, MN, USA, 2005).
2. Bertin, J. *Sémiologie graphique : Les diagrammes - Les réseaux - Les cartes*. Editions de l'Ecole des Hautes Etudes en Sciences, Paris, France, 1967.
3. Climer, S. and Zhang, W., Take a Walk and Cluster Genes: a TSP-Based Approach to Optimal Rearrangement Clustering. in *21st International Conference on Machine Learning* (Banff, Alberta, Canada, 2004).
4. Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. Cluster Analysis and Display of Genome-Wide Expression Patterns. *Proc. Natl. Acad. Sci. USA*, *95*. 14863-14868.
5. Rao, R. and Card, S.K. The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information in *Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence* ACM Press, Boston, Massachusetts, United States 1994 318-322
6. Schneiderman, B., The Eyes Have It: A Task by Data Type Taxonomy for Information Visualization. in *IEEE Symposium on Visual Languages*, (1996), p 336.
7. Spenke, M., Beilken, C. and Berlage, T. FOCUS: the interactive table for product comparison and selection in *Proceedings of the 9th annual ACM symposium on User interface software and technology* ACM Press, Seattle, Washington, United States 1996 41-50
8. Wehrend, S. and Clayton, L., A Problem-Oriented Classification of Visualization Techniques. in *IEEE Conference on Visualization*, (1990), pp 139-143.