Investigating and Reflecting on the Integration of Automatic Data Analysis and Visualization in Knowledge Discovery

Enrico Bertini
Université de Fribourg
Bd de Pérolles 90
Fribourg, Switzerland
enrico.bertini@unifr.ch

Denis Lalanne
Université de Fribourg
Bd de Pérolles 90
Fribourg, Switzerland
denis.lalanne@unifr.ch

ABSTRACT

The aim of this work is to survey and reflect on the various ways visualization and data mining can be integrated to achieve effective knowledge discovery by involving the best of human and machine capabilities. Following a bottom-up bibliographic research approach, the article categorizes the observed techniques in classes, highlighting current trends, gaps, and potential future directions for research. In particular it looks at strengths and weaknesses of information visualization (infovis) and data mining, and for which purposes researchers in infovis use data mining techniques and reversely how researchers in data mining employ infovis techniques. The article then proposes, on the basis of the extracted patterns, a series of potential extensions not found in literature. Finally, we use this information to analyze the discovery process by comparing the analysis steps from the perspective of information visualization and data mining. The comparison brings to light new perspectives on how mining and visualization can best employ human and machine strengths. This activity leads to a series of reflections and research questions that can help to further advance the science of visual analytics.

Keywords

Visualization, Data Mining, Visual Data Mining, Knowledge Discovery, Visual Analytics.

1. INTRODUCTION

While information visualization (infovis) targets the visual representation of large-scale data collections to help people understand and analyze information, data mining, on the other hand, aims at extracting hidden patterns and models from data, automatically or semi-automatically.

In its most extreme representation, infovis can be seen as a human-centered approach to knowledge discovery, whereas data mining is generally purely machine-driven, using computational tools to extract automatically models or patterns out of data, to devise information and ultimately knowledge.

Interactive Machine Learning (or Interactive Data Mining) [1][2] is an area of research where the integration of human and machine capabilities is advocated, beyond the scope of visual data analysis, as a way to build better computational models out of data. It suggests and promotes an approach where the user can interactively influence the decisions taken by learning algorithms and make refinements where needed.

Visual analytics is a new interdisciplinary domain that integrates several domains like: interactive visualization, statistics and data mining, human factors, to focus on analytical reasoning facilitated by interactive visual interfaces [3]. Often, it is presented as the combination of infovis techniques with data mining capabilities to make it more powerful and interactive. According to Keim et al., visual analytics is more than just visualization and can rather be seen as an integrated approach combining visualization, human factors and data analysis [4].

At the time of writing, we realize that regardless several efforts exist to define what visual analytics is on a higher level (above all the Visual Analytics Agenda [3]), we still lack a detailed analysis of: 1) how currently the existing techniques integrate and to what extent; 2) what other kinds of integrations might be achieved.

The purpose of this work is to start shedding some light on these issues. To this end, we have performed a literature review of papers from premier conferences in data mining and information visualization, extracting those in which some form of integration exists. The analysis permitted to categorize the observed techniques in classes. For each class we provide a description of the main observed patterns, and then we discuss potential extensions we deem feasible and important to realize. The analysis follows by a comparison of the analytical processes as they happen in data mining and in visualization. This comparison, together with the knowledge gained in the literature review, permits to clarify some commonalities and differences between the automatic and visual approaches. We believe this kind of reasoning can help framing the problem of automatic and interactive analysis and better understand the role of the human and the machine.

Given the nature of our discussion we might seem to suggest that visualization and data mining are always two competing methods to address the same problem and that some form of integration is always desirable. But, this is not the message we want to convey here. Rather we want to draw a picture of what can happen, and should happen, when we focus our attention on only those cases where an overlap exists and integration is desirable; without discussing in any details when this is desirable.

The paper is organized as follows. Section 2 introduces some terminology to clarify the meaning of some words that often appear when talking about automatic or interactive data analysis. Section 3 introduces the literature review and its methodology. Section 4 illustrates the result of the review describing the patterns we found. Section 5 describes the extensions we propose. Section 6 dissects commonalities and differences between the

analysis processes and provides some reflections on further research in data mining and visualization. Section 7 elaborates on the reflections and introduces the idea of defining visual analytics problems. Finally, Section 8 discusses the limitations of this work, and thus provides ideas for its future extension, and Section 9 closes the paper with conclusions.

2. TERMINOLOGY

The common goal of information visualization (infovis) and data mining is to extract knowledge from raw data, through visualization techniques and automatic computational analysis respectively. In the rest of this article, we both use the terms infovis and visualization when speaking about the first approach, and indifferently about data mining or automatic data analysis when speaking about the second. Before going further in our inspection of the integration of the two approaches, we thought useful to agree on the definition of basic concepts such as data, information, knowledge, model, pattern and hypothesis and on how they are linked in the knowledge discovery process. Some definitions below are inspired from a mix of sources (e.g. www.infovis-wiki.net, Oxford English Dictionary) and from our own thoughts. The way they relate in the knowledge discovery process is our own interpretation and as such can be further discussed.

In the context of *knowledge discovery*, raw data are the lowest level of abstraction; *data* refers to a collection of facts usually collected by observations, measures or experiments. It is called abstract data in infovis, since it refers to data that have no inherent spatial structure enabling further mapping to any geometry. From data, models and patterns can be extracted, either automatically using data mining techniques or by humans using their conceptual, perceptual or visual skills respectively. The use of human intuition to come up with observations about the data is generally called *insight*, i.e., the act or outcome of grasping the inward or of perceiving in an intuitive manner.

Patterns and models are not necessarily linked, even though some authors consider them as synonyms. A *pattern* is made of recurring events or objects that repeat in a predictable manner. A *model* is a mathematical representation of a system phenomena, or processes. It is basically a simplified abstract view of the complex reality. One way to distinguish models and patterns is the following: patterns are directly attached to data or a sub-set of data; whereas models are more conceptual and are extra information that cannot necessarily be observed visually in the data. Further, the observation of some patterns can result in a model and inversely, the simulation of a model can result in a pattern.

Hypotheses are human artifacts and are derived from models and patterns. A *hypothesis* consists either of a suggested explanation for an observable phenomenon or of a reasoned proposal predicting a possible causal correlation among multiple phenomena. A validated hypothesis becomes information that can be communicated. Finally, information reaches the solid state of knowledge when it is crystallized, i.e., it reaches the most compact description possible for a set of data relative to some task without removing information critical to its execution.

3. LITERATURE REVIEW

We started our analysis with a literature review in order to ground our reasoning on observed facts and limit the degree of subjectivity. We followed a mixed approach in which bottom-up and top-down analyses have been mixed to let the data speak for themselves and suggest new ideas or use the literature to investigate our assumptions or formulated hypotheses.

We included in the literature papers from major conferences in information visualization, data mining, knowledge discovery and visual analytics. In the current state of our analysis the papers have been selected from the whole set of proceedings of: *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, *IEEE International Conference on Data Mining (ICDM)* and the *IEEE Symposium on Information Visualization (InfoVis)*. We selected infovis candidate papers searching in the IEEE Explore[1] library using keywords like: "data mining", "clustering", "classification", etc. Reversely, in data mining conferences we looked for keywords like: "visualization", "interaction", etc. Manual skimming followed paper extraction to remove spurious papers. The final set of retained papers counts 48 items. Table 1 shows the distribution of the retained papers according to the paper source and the classification of papers presented below.

SOURCE	NUM. OF PAPERS	V++	M++	VM
KDD ('95-'08)	20	7	9	4
ICDM ('01-08)	14	5	5	4
INFOVIS ('95-'08)	14	9	5	0

Table 1 - Distribution of the final list of retained papers according to source (conference) and paper type.

The whole list of reviewed papers with attached notes and categories can be found at the following address: http://diuf.unifr.ch/people/bertinie/ivdm-review.

4. PAPER CATEGORIES AND OBSERVED PATTERNS

We used various dimensions in order to classify the chosen papers: the knowledge discovery step it supports, whether it is interactive or not, the major mining and visualization techniques used, etc. In particular, in regards to the aim of this paper, we classified the paper according to four major categories indicating which approach drives the research:

[1] http://ieeexplore.ieee.org/

- **Computationally enhanced Visualization (V++)** contains techniques which are fundamentally visual but contain some form of automatic computation to support the visualization;
- **Visually enhanced Mining (M++)** contains techniques in which automatic data mining algorithms are the primary data analysis means and visualization provides support in understanding and validating the result;
- **Integrated Visualization and Mining (VM)** contains techniques in which visualization and mining are integrated in a way that it's not possible to distinguish a predominant role of any of the two in the process.

Since the focus of this paper is on how visualization and mining can cooperate in knowledge discovery, in the paper we do not discuss other categories we have built during the process where a very predominant role of mining or visualization was present. More specifically, we did not take into account the pure visualization category containing techniques based exclusively on visualization, without any type of algorithmic support, or the pure mining category, making no use of visualization techniques.

4.1 Enhanced Visualization (V++)

This category pertains to techniques in which *visualization* is the primary data analysis means and automatic computation (that is the "++" in the name) provides additional features to make the tool more effective. In other words, when the "++" part is removed it becomes a "pure" visualization technique.

The techniques collected in our literature review are organized around three main patterns (Projection, Data Reduction, Pattern Disclosure) that represent different benefits brought by automatic computation to the information visualization process.

Projection. Automatic analysis methods often take place in the inner workings of visualization, by creating a mapping between data items and their graphical objects' position on the screen. They all share the idea that the position assumed by a data point on the screen is not the result of a direct and fixed mapping rule between some data dimensions and screen coordinates, but rather of a more complex computation that takes into account all data dimensions and cases. Ward refers to this kind of placement techniques as "Derived Data Placement Strategies" in his glyph placement taxonomy [5]. The most traditional technique in this class is Multidimensional Scaling (MDS). Figure 1 shows an example of MDS taken from [6], where a fast MDS visualization algorithm is proposed.

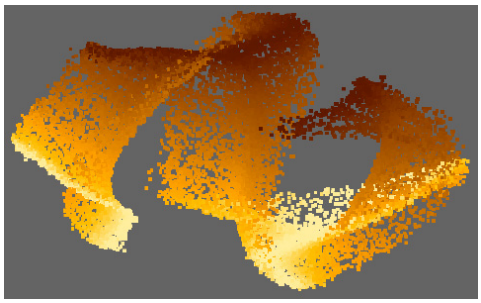

Figure 1 - Multidimensional Scaling. Example of projection technique (extracted from [6]).

But in literature it is possible to find many variations and alternatives like graph drawing algorithms and other complex spatialization techniques.

Intelligent Data Reduction. Data reduction is another area where computation can support visualization. Visualization has very well known scalability problems that limit the number of data cases or dimensions that can be shown at once. Automatic methods can reduce data complexity, with controlled information loss, and at the same time allow for a more efficient use of screen space. Pattern matching techniques can replace data overviews with visualizations of selected data cases that match a user-defined query. Sampling can reduce the number of data cases with controlled information loss. Feature selection can reduce data dimensionality by retaining subsets that carry the large majority of useful information contained in the data (and thus are most likely to show interesting patterns). Figure 2 shows an example of dimension filtering as proposed in [7]

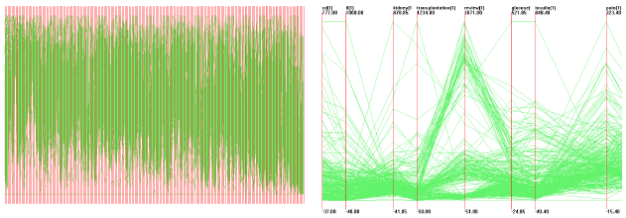

Figure 2 - Dimension filtering. Example of data reduction technique (extracted from [7]).

Patterns Disclosure. In several visualization techniques the effectiveness with which useful patterns can be extracted depends on how the visualization is configured. Automatic methods can help configure the visualization in a way that significant features pop-out from the screen. Axes-reordering in parallel coordinates is one instance of such case [8]. Similarly, in visualizations where the degree of freedom in visual configuration is limited, pattern detection algorithms can help make visual patterns more prominent and thus readily visible. For instance, Vizster [9] (in Figure 3) organizes the nodes of a social network graph around automatically detected clusters enclosed within colored areas. Johansson et al. in [10] describe an enhanced version of Parallel Coordinates where clustering and a series of user-controlled transfer functions help the user reveal complex structures that would be hard, if not impossible, to capture otherwise.

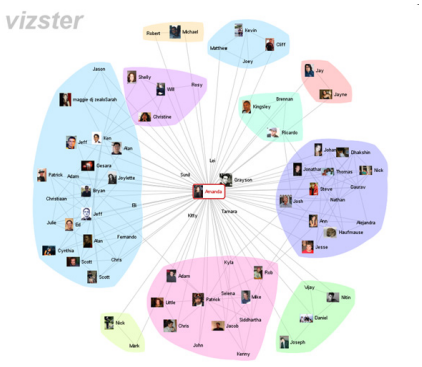

Figure 3 - Graph clustering in Vizster. Example of pattern disclosure technique (extracted from [9]).

4.2 Enhanced Mining (M++)

This category pertains to techniques in which *data mining* is the primary data analysis means and visualization (that is the "++" in the name) provides an advanced interactive interface to present the results. In other words, when the "++" part is removed it becomes a "pure" data mining technique.

The techniques collected in our literature review can be organized around two major patterns (Model Presentation and Pattern Exploration & Filtering) that represent different benefits brought by visualization to data mining.

Model Presentation. Visualization is used to facilitate the interpretation of the model extracted by the mining technique. According to the method used, the ease with which the model is interpreted can vary. Some models naturally lend themselves to visual abstraction (e.g., dendrogram in hierarchical clustering) whereas some others require more sophisticated designs (e.g., neural networks or support vector machines) because a natural metaphor simply does not exist. One example of model visualization is "Nomograms" [11] (Figure 4), where the output of a classification or regression algorithm (e.g., SVM) is presented in a way to understand the relationship between dimensions and target variable.

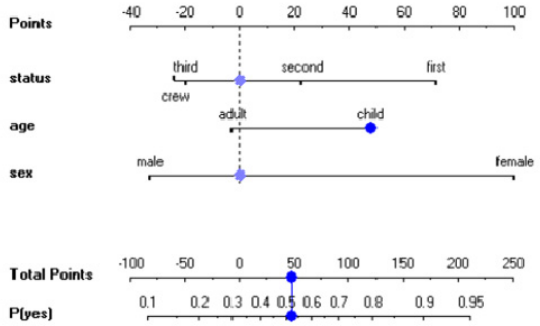

Figure 4 – Nomogram visualization. Example of model presentation technique (extracted from [11]).

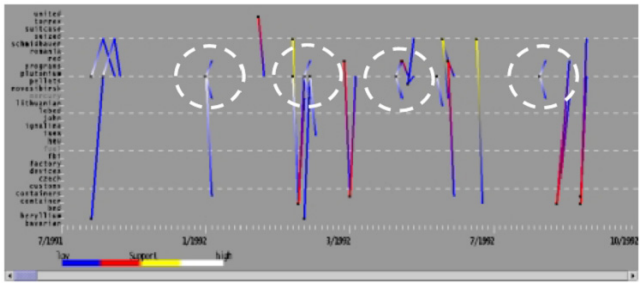

Figure 5 - Sequential Patterns visualization. Example of patterns exploration and filtering (extracted from [12]).

Beyond model interpretation, visualization also works as a way to visually convey the level of trust a user can assign to the model or parts of it. Interactions associated to the visualization permits to "play" with the model allowing for a deeper understanding of the model and its underlying data.

Patterns Exploration and Filtering. Some mining methods generate, in place of descriptive or predictive models, complex and numerous patterns which are difficult to summarize in a compact representation (e.g., association rules). In this case, visualization often adopts techniques similar to plain data visualization and the patterns are managed like raw data. Visualization here helps gaining an overview of the distribution of these patterns and making sense of their nature. Interactive filtering and direct manipulation tools have a prominent role, in that finding the interesting pattern out of numerous uninteresting ones is the key goal. An example is the output obtained from sequential temporal patterns, as shown in Figure 5, where numerous time patterns are presented to understand how topics change in a stream of news [12].

4.3 Integrated Visualization & Mining (VM)

This category combines visualization and mining approaches. None of them predominate over the other and ideally they are combined in a synergic way. In the literature we found two kinds of integration strategies that we describe below. The two approaches described below illustrate the two extremes to integrate mining and visualization.

White-Box Integration. In this kind of integration the human and the machine cooperate *during* the model building process in a way that intermediary decisions in the algorithm can be taken either by the user or the machine. This kind of systems is quite rare. There are examples of cooperative construction of classification trees, like the one presented in [13], where the user steers the construction process and at any stage can ask the computer to make one step in his or her place like splitting a node or expanding a sub-tree. These systems show the highest degree of collaboration between the user and the machine and go beyond the creation of accurate models. They help building trust and understanding, because the whole process is visible, and also they permit to directly exploit the user's domain knowledge in the model construction process. One notable example of this process is the one described in [13] (Figure 6), where the intermediary steps of a decision tree algorithm can be taken interactively or automatically.

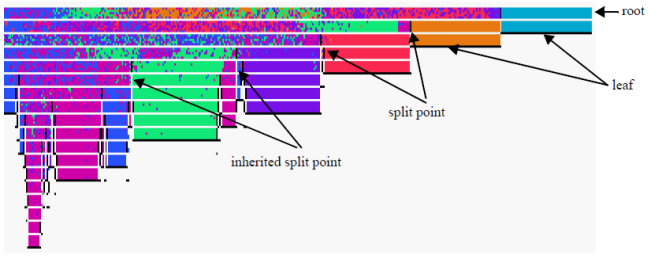

Figure 6 - Collaborative decision tree construction. Example of white-box integration (extracted from [13]).

Black-Box Integration (feedback loop). Integration between mining and visualization can also happen indirectly using the algorithm as a black box, but giving the user the possibility to "play" with parameters in a tight visual loop environment, where changes in the parameters are automatically reflected in the

visualization. In this way the connection between parameters and model, even if not explicit, can be intuitively understood. Alternatively, the same integration can be obtained in a sort of "relevance feedback" fashion, where the system generates a set of alternative solutions and the user instructs the system on which are the most interesting ones and gives hints on how to generate a new set. An example of this kind of integration is in [14] where "bracketing" is used to show alternative solutions of a subspace clustering algorithm simultaneously (Figure 7).

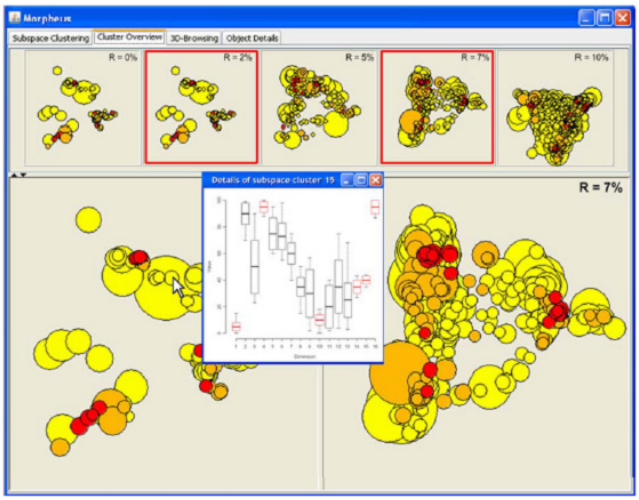

Figure 7 - Bracketing technique is subspace clustering. Example of black-box techniques (extracted from [14]).

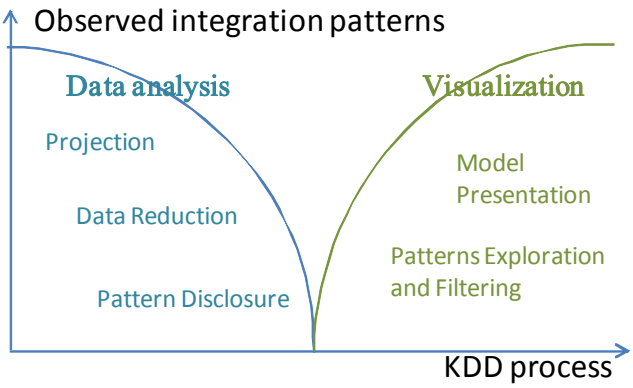

Figure 8 - In current enhanced infovis systems (V++), data analysis is mostly used at the beginning of the KDD process and reversely, in enhanced data analysis systems (M++), visualization is principally used at the end.

Table 2, at the end of Section 5, summarizes the observed integrations of visualization and data analysis in data mining systems, as described above, and suggested novel integrations as described in the next section.

5. SUGGESTED INTEGRATIONS

The previous section listed the major traits we found in current visual analytics systems, through our literature survey. As presented in the previous section and illustrated in Figure 8, the integration patterns in enhanced visualization or enhanced mining systems is stereotyped. On one hand, data analysis is used by infovis practitioners at the beginning of the KDD process to project and reduce data, and disclose patterns. Reversely, visualization is mostly used by data analysis practitioners at the end of the process to present visually a model or to explore and filter the patterns they find.

In the following we propose to enhance the respective contributions of data analysis and visualization to better cover the full KDD spectrum, towards a tighter integration.

5.1 Enhancing data analysis contribution

All the automatic data analysis methods described in section 4.1 share the common goal of helping the user more easily extract information from the visualization. But, if we take into account the broader picture of data analysis and analytical reasoning, we see that automatic techniques could also be employed to go beyond simple pattern detection, and intervene at later stages of the knowledge discovery process. Below we list some of the functions that we believe would be beneficial to information visualization.

Visual Model Building. One limitation of current visualization systems is their inability to go beyond simple pattern detection and frame the problem around a schema to enable higher level reasoning and hypothesis generation. Ideally, the user should be able to find connections among the extracted patterns to build higher level hypotheses and complex models. This is another area where data mining has an advantage over visualization in that in the large majority of the existing methods a specific conceptual model is inherent in the technique. *Classification* and *regression*, for instance, imply a functional model: an instantiation of the set of predictive variables produces a target value. *Clustering* implies a grouping model, where data is aggregated in groups of items that share similar properties. *Rules* imply an inductive model where if-then associations are used. This kind of mental scaffold is usually absent in visualization, or better it is formed only in the user's mind. But there's no inherent reason why future systems might not be provided with visual modeling tools that permit, on the one hand to keep the level of flexibility of visualization tools, on the other hand to structure the visualization around a specific model building paradigm. Two rare examples of systems that go towards this direction are PaintingClass [15] and the Perception Bases Classification (PBC) system [16] in which classification can be carried out interactively by means of purely visual systems.

Verification and Refinement. One notable feature of automatic data mining over data visualization is its ability to communicate not only patterns and models but also the level of trust a user can assign to the extracted knowledge. Similar functions are usually not present in standard visualization tools and surprisingly little research as been carried out towards this direction so far. Automatic algorithms could be run on extracted patterns to help the user assess their quality once they are detected. To date, the only systems we are aware of where a similar idea has been implemented are [17][18], where respectively data abstraction quality is measured and progressive automatic refinement of visual clusters is performed.

Another related area of investigation is the use of the traditional split in *training data* and *test data* used in supervised learning as a novel paradigm to use in data visualization. There is no reason in principle not to use the same technique in information visualization to allow for verification of extracted patterns. Some

few studies on sampling for data visualization slightly touch on this issue [19][20] but none of them focuses on the use of sampling or data segmentation for verification purposes.

Prediction. Worthy of special remark is also the almost complete absence of predictive modeling in visualization, as highlighted by Amar and Stasko in their analysis of "analytic gaps" in information visualization [21]. While it is fairly simple to isolate data segments and spot correlations, even in multidimensional spaces, current information visualization tools lack the right affordances and interactive tools to structure a problem around prediction. Questions like: "which data dimensions have the highest predictive power?", "what combination of data values are needed to obtain a target result?" are not commonly in the scope of traditional visualization tools. Many real world problems are based on prediction, like the ones involved in marketing campaigns or financial projections, and there is no reason to believe that visualization cannot play a far larger role in this domain.

5.2 Enhancing visualization contribution

Visualization applied to data mining output, as shown in section 4.2, provides great benefits in terms of model interpretation and trust-building. We believe that visualization, however, can provide additional benefits that have not been fully exploited so far, and enable users to intervene in earlier stages of the knowledge discovery process.

Visualizing the Parameter Space and Alternatives. One of the characteristic features of data mining is its capability of generating different results and models by manipulating a limited set of parameters. This is common to all methods and can be seen as both an advantage and a limitation. It is an advantage in that the necessary flexibility is given to create alternatives and adapt to different analytic goals. But, it is also a big limitation in terms of interaction, in that setting the parameters of a mining algorithm is often perceived by the user as an "esoteric" activity in which the relation between actions and results is not evident. Even more problematic, when alternative models are constructed, is extremely complicated to compare them in the space of a single user interface. Visualization in our opinion has the power to bridge these gaps by: 1) providing means to more directly represent the connection between parameters and results; 2) allowing for visualization structures that permit the comparison of alternative results. Parameter space visualization is, to the best of our knowledge, a totally unexplored and yet extremely needed research area. Ideally, by visualizing the parameter space it would be possible not only to understand the connection between parameter values and outcome but also to explore the "sensitivity" of certain parameters and their interaction. Comparison of alternative results is also related and interesting in that visualization has the power to provide the right tools to compare alternative visual abstractions; as demonstrated for instance by the success of the systems presented at the InfoVis 2003 contest on Pair Wise Comparison of Trees [22]. One system in our literature review partially supports this kind of comparison by generating different alternative results of a subspace clustering algorithm [14]. The user can see the results obtained through the variation of various parameters and choose the most interesting ones among the set of available results. But, unfortunately, the concept is not researched in depth or further generalized.

Model-Data Linking. The models that mining algorithms create out of data are higher level data abstractions that permits to summarize complex relations out of large data. If from the one hand these abstractions facilitate data analysis and reduce the complexity of the original problem space, from the other hand the abstraction process creates a semantic gap. The abstractions often make it difficult to interpret the observed relations in terms of the original data space and the observed objects in terms of the application area. Most systems in our literature survey provide model representation, but very rarely they permit to drill down to the data level to link an observed relation to its underlying data. In some cases such a lack of connection between model and data can create relevant limitations in model understanding and trust building, and visualization is the right tool to bridge this gap. One example is data clustering. Besides the large provision of visual and interactive techniques to represent clustering results, it is very rare to find systems where the linkage between extracted clusters and data instances is made explicit by the visualization. And this is somewhat surprising in that the goal of data clustering is not only to partition data in a set of homogeneous groups but also, and potentially more important, to characterize them in a way that their content can be described in terms of few data dimensions and values. A better connection between model and raw data is then useful also to spot relevant outliers, which can often triggers new analyses and lines of thought. Without such a capability the analyst is forced to base his reasoning only on abstractions, thus limiting the opportunities for serendipitous discoveries and trust building.

One notable example where such connection is implemented is the Hierarchical Clustering Explorer [23], where at each time the user can easily drill down form the clustering tree up to a single data item in the original data table.

	Observed	Suggested
V++	Projection Intelligent Data reduction Pattern Disclosure	Visual Model Building Verification and Refinement Prediction
M++	Model Presentation Patterns Exploration & Filtering	Visualizing Parameter Space & Alternatives Model-Data Linking
VM	White-Box Integration Black-Box Integration	Mixed Initiative KDD

Table 2 – Summary of observed and suggested integrations of visualization and data analysis in visual mining systems.

5.3 Towards a mixed-initiative KDD process

Having analyzed a wide spectrum of integrations between automatic and interactive methods as summarized in Table 2, we believe that one of the most interesting and promising direction for future research is to achieve a full mixed-initiative KDD process where the human and the machine cooperate at the same level. As shown in Figure 9, with the suggested contributions presented in this article, visualization and data analysis, and as such humans and machines respectively, can both contribute throughout the whole KDD process.

Humans and machines are complementary, and visualization and data mining should make use of the specificities of each. Humans

are intuitive and have remarkable interpretation skills involving the analysis context and accumulated domain knowledge. They are good at getting the "big picture" and at performing high level reasoning towards knowledge. Machines on the other hand are fast and reliable at computing data, and are less prone to errors.

While humans are good at choosing modeling strategies through visualization, the machine is good at computing large amounts of data for projecting and reducing data. Machines can disclose and highlight all the patterns found automatically over data, humans can assign a meaning to them and keep only the most interesting ones, according to their knowledge of the data and its domain. Furthermore, human and machine can collaborate to build models, either coming from mining models or alternatively derived by humans through their perceptive and cognitive systems. At this stage visualization techniques can be particularly useful to bridge the gap between data and the extracted models. Finally, data mining techniques can be useful to support the validation of observed model or knowledge that humans can ultimately refine through interaction.

To date, the only system that comes closer to the idea of a mixed-initiative KDD process is the one we mentioned above in White-Box Integration [13], where a decision tree can be constructed by alternating steps of human-based decisions and machine-based algorithmic steps.

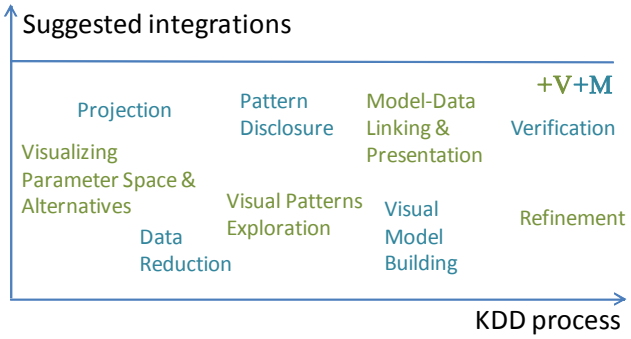

Figure 9 – With the suggested contributions presented in this article, visualization and data analysis both contribute throughout the whole KDD process.

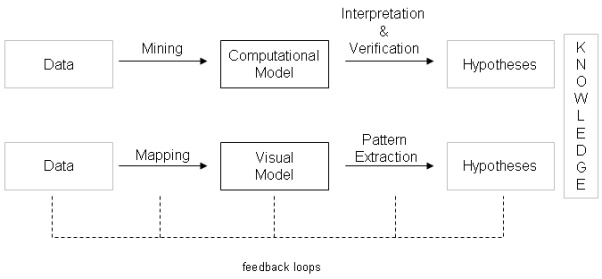

Figure 10 - Comparing mining and visualization analytic processes.

6. REFLECTIONS

Visualization and data mining are currently alternative methods to transform data into knowledge. Having said that, a legitimate question stands: are they just different recipes to cope with the same problems or do they differ in any substantial manner? We believe that answering this question is becoming of increasing importance as we attempt to get the most out of the two disciplines and create successful integrations like the one advocated in Visual Analytics.

In Figure 10 we provide an extremely simplified model of the visualization (bottom) and mining (top) processes to put into relation the steps that transform data into knowledge.

We have two main set of processes that we deem important to compare: from *data to models* and from *models to hypotheses* generation.

From data to models

In mining, data are transformed into a *computational model* through a *mining process*, whereas in visualization they are transformed into a *visual model* through a visual *mapping process*. The comparison of these two steps can lead to interesting questions and insights. For instance, both mining and mapping require the definition of a schema (visual or functional) around which data is modeled. The definition of such a schema has a strong relationship with the mental model the users have because it influences, either explicitly or implicitly, the perspective they use in understanding the data. In visualization such a schema is notoriously flexible and can easily lead to the exploration of different views on data. Such flexibility is not common in automatic systems, where parameter setting is a quite cumbersome activity. Therefore one research question to explore is: "*how can we transfer such flexibility into the mining process so that it becomes easier to explore alternative solutions?*" On the other side, the mining process is notoriously robust and equipped with reliable methods to verify the quality and trustworthiness of the outcome. Therefore another pertinent question is: "*how can we transfer such robustness and verification capabilities into visualization?*" The integrations suggested in the previous section point to some potential solutions, but it is evident that there is a whole space of possibilities to explore.

From models to hypotheses

Continuing on our comparison between the two processes, we see that in mining the main user mental activity involved is the *interpretation* of the extracted model, whereas in visualization the main user's activity concerns the *visual extraction* of data patterns from the screen.

The output of a mining algorithm is some kind of formal abstraction of data; therefore it is necessary to provide an interface to understand the model, its relation to data, and its validity. Here its worth to point out that traditional data mining is often presented as simply not having an interface, but this is hardly true. Mining systems are not without an interface, they just provide simple and minimalistic interfaces, like results organized in a tabular fashion. The question therefore is not necessarily how to "attach" and interface to a mining system, but rather how to make it more effective through visualization.

From the visualization design point of view it is important to recognize the shift of mental activity from understanding and interpretation of a computational model to the extraction of visual patterns. One main question here is: "*how can we provide effective visualizations to interpret, understand, and verify computational models?*" Even if it is reasonable to believe that

what we learned from *data visualization* will be easily applied to *model visualization* design, it's important to recognize that model visualization is far less developed and that new requirements or design challenges may emerge. For this reason, another relevant research question is: "*is model visualization fundamentally different from data visualization in terms of visual metaphors, interaction techniques, and design solutions?*"

Finally, as we have noted in the previous sections, the pattern extraction activity in the visualization process can be aided by automatic procedures as those found in the mining process. One last research question is therefore: "*how can automatic data analysis support users more easily extract relevant and accurate patterns out of data?*"

6.1 Interaction: The feedback loop

So far, we have only discussed one direction of the human-machine interface, that is, from the machine to the human. The opposite direction is often neglected but it is equally important because it permits to close the interaction feedback loop. It is in fact the possibility to iterate over alternate phases of human perception and understanding of the current state and human actions to change state and devise alternatives that fuel the discovery and learning process both in mining and visualization.

On a higher level this is also how the Sensemaking Theory describes how people make sense of information. As Pirolli and Card note in [24], the sense making process revolves around "one set of activities that cycle around finding information and another that cycles around making sense of the information".

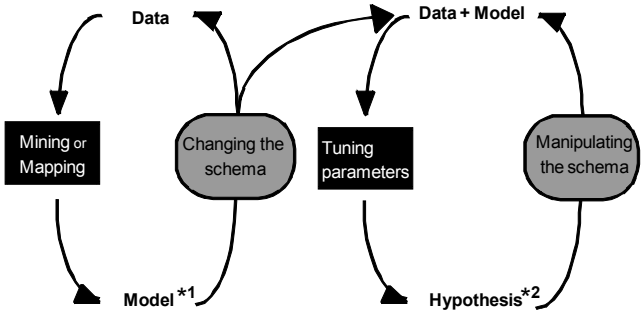

Figure 11 – The feedback loop in Knowledge Discovery. The grey boxes represent the two major stages at which humans can intervene. (*1) A model can be either computational or visual. (*2) Users tune parameters of the visualization or of the computational model, until they confirm their hypothesis.

In our literature review, almost half of the papers do not propose or describe means to interact with the system and as such to intervene on the knowledge discovery process. In the 55 papers reviewed, the major interaction techniques found can be grouped in two categories depending on the knowledge discovery step at which users can intervene, i.e., pre or post model creation, to change the schema or manipulate it respectively as illustrated on Figure 5.

Changing the schema. Both in visualization and in data mining at any stage the user can decide to change the schema. In visualization changing the schema means changing the visual mapping in a way that data can be seen under a new perspective. In data mining it means reframing the problem so that it is represented under a new model; as when, for instance, moving the analysis from the generation of rules to finding data clusters. This kind of activities is often neglected and yet it is very important because as the user's mental model changes the tools must adapt in a way to reflect this change. The goodness of a data analysis system should be measured also in terms of this flexibility. This need of reframing problems under different schemes uncover a relevant gap in current tools. One of the biggest challenges in visualization is to find an appropriate visualization for the task at hand. Despite numerous efforts towards this direction, especially at the early stage of information visualization (e.g., in Jock MacKinlay's work [25]), current tools offer very limited support. Automatic or semi-automatic methods should be employed to help users find appropriate visual mappings or yet suggest possible alternatives.

Manipulating and tuning the schema. Another user's option to create alternative views or models is to change parameters within the context of a given schema. In visualization this is normally achieved by manipulating a view though interactions like: dynamic filtering, axes reordering, zoom & pan, etc. In data mining it involves some form of parameter tuning, as when using different distance functions or number of desired groups in data clustering. This last function is of special interest in that visualization can be a powerful means to help users tune up their mining models. As we have already discussed in Section 5 in "Visualizing Alternatives", the use of powerful visualization and interaction schemes could greatly improve the state of current tools. Of special interest is the study of efficient techniques that permit to understand how a model changes when one or more parameters change. In current tools it is almost impossible to achieve this level of interaction. Not only the large majority of parameters are difficult to interpret but also the user is forced to go through a series of "blind" trial-and-error steps where the user changes some parameters, waits for the construction of the new model, evaluates the result and iterates over until he or she is satisfied.

7. DEFINING VISUAL ANALYTICS PROBLEMS

The work we have described here stems from the assumption that the fingerprint of Visual Analytics is the integration of automatic and interactive data analysis. Our belief is that this approach has the potential of not only uncovering new research directions but also of defining the very nature of Visual Analytics.

Regarding this last point, however, we acknowledge that a whole set of new research approaches are needed. Instead of surveying the complementarities between data mining and infovis techniques, like we did in this survey, another useful approach would consist in studying the type of problems or applications which are best addressed by data mining or by visualization alone. In other words, not all problems are Visual Analytics problems and it is necessary to understand when Visual Analytics should **not** be used, as pointed out by Daniel Keim in [26].

Therefore our community is confronted with tough questions like: "*Are the problems addressed by data mining of a different class than problems addressed by infovis?*" and "*If they are different,*

why are they different?" or "*When it is not advisable to use a Visual Analytics solution to a given problem?*"

Standard data mining problems are generally clearly defined, e.g. categorize data, find clusters, etc. Infovis techniques on the other hand support exploration and communication. The tasks supported by visualization are clearly open-ended and cannot be reduced to a single problem solving task. Rather than supporting problem solving, visual analytics systems could rather support practitioners in understanding the nature of the data and in better defining the problem to be solved.

We suggest that in the future, the community tries to categorize problems for which data mining is perfectly suited, and reversely problems for which using visualization, and thus involving humans, is mandatory. This could, hopefully, lead towards the definition of new contests in which both infovis and data mining practitioners can compete.

8. LIMITATIONS AND FUTURE WORK

Despite our effort to produce a meaningful literature survey and to extract useful indication out of it, we believe it is important to highlight and acknowledge some limitations of this work.

The literature we have analyzed, though useful, is far from being a full survey. We decided to use a number of papers that could be analyzed in a relative short time by the two authors. It can be considered a large enough sample to draw meaningful trends, explore potential extensions, and highlight pertinent research questions.

As a consequence we decided not to draw any statistics out of our study. The literature contains some hand-made categorizations that could have been used to further categorize the techniques and depict some additional trends out of it. We postpone this task to later works.

Finally, it's important to take into account that a large part of this paper is the product of subjective indications stemming from what we believed worth to extract from the literature. Nonetheless, we believe that our analysis and guidelines can highlight hidden patterns and stimulate further research on important issues in this cross-disciplinary topic.

We plan to advance this work after having received sufficient feedback from the community. We want to explore in more details the problem of better defining Visual Analytics. In particular, we want to investigate in more depth the characterization of visual analytics problems and understand what differentiates a visual analytics problem from other types of problems.

9. CONCLUSIONS

We have presented a literature review on the role of visualization and data mining in the knowledge discovery process. From the review we have generated a series of classes through which we have categorized the collected papers: the knowledge discovery step it supports, whether it is interactive or not, the major mining and visualization techniques used, etc. In particular, in regards to the aim of this paper, we classified the paper according to three major categories indicating which approach drives the knowledge discovery: computationally enhanced visualization systems, visually enhanced data mining systems, and integrated visual and mining systems.

This categorization highlights some observed patterns and suggests potential extensions which are not present in the considered literature. For instance, in order to enhance the standard visualization process, we believe data mining techniques could support visual model building to go beyond simple pattern detection. Further, mining techniques could be also used to verify and assess the quality of patterns detected by users. Reversely, visualization could enhance the data mining process to visualize modeling alternatives, and to understand modeling results through a better model-data linking and presentation.

In addition to these suggestions, the article provides a series of higher level reflections on the analysis process as it happens in visualization and data mining. These reflections suggest new perspective on the role of visualization and mining in the data analysis process and potential areas of investigation towards a better integration of both. In particular, this study suggests improving the human machine interaction through a better consideration of the feedback loop so that users can intervene at different levels of the knowledge discovery process, to change and manipulate the schema respectively.

REFERENCES

[1] J.A. Fails and J. Olsen, "Interactive machine learning," *IUI '03: Proceedings of the 8th international conference on Intelligent user interfaces*, New York, NY, USA: ACM, 2003, pp. 39–45.

[2] M. Ware, E. Frank, G. Holmes, M. Hall, and I.H. Witten, "Interactive machine learning: letting users build classifiers," *International Journal of Human Computer Studies*, vol. 55, 2001, pp. 281–292.

[3] J.J. Thomas and K.A. Cook, *Illuminating the path: The research and development agenda for visual analytics*, IEEE, 2005.

[4] D.A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler, "Visual analytics: Scope and challenges," *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*, Springer, 2008, pp. 76–90.

[5] M.O. Ward, "A taxonomy of glyph placement strategies for multidimensional data visualization," *Information Visualization*, vol. 1, 2002, pp. 194–210.

[6] A. Morrison, G. Ross, and M. Chalmers, "Fast multidimensional scaling through sampling, springs and interpolation," *Information Visualization*, vol. 2, 2003, pp. 68-77.

[7] P. Yang, "Interactive Hierarchical Dimension Ordering, Spacing and Filtering for Exploration of High Dimensional Datasets," Oct. 2003.

[8] W. Peng, M.O. Ward, and E.A. Rundensteiner, "Clutter reduction in multi-dimensional data visualization using dimension reordering," *IEEE Symposium on Information Visualization, 2004. INFOVIS 2004*, pp. 89–96.

[9] J. Heer and D. Boyd, "Vizster: Visualizing online social networks," *Proceedings of the 2005 IEEE Symposium on Information Visualization*, 2005, pp. 33–40.

[10] J. Johansson, P. Ljung, M. Jern, and M. Cooper, "Revealing Structure within Clustered Parallel Coordinates Displays," *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*, IEEE Computer Society, 2005, p. 17.

[11] A. Jakulin, M. Možina, J. Demšar, I. Bratko, and B. Zupan, "Nomograms for visualizing support vector machines," *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, Chicago, Illinois, USA: ACM, 2005, pp. 108-117.

[12] Pak Chung Wong, W. Cowley, H. Foote, E. Jurrus, and J. Thomas, "Visualizing sequential patterns for text mining," *Information Visualization, 2000. InfoVis 2000. IEEE Symposium on*, 2000, pp. 105-111.

[13] M. Ankerst, M. Ester, and H. Kriegel, "Towards an effective cooperation of the user and the computer for classification," *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2000, pp. 179-188.

[14] E. Müller, I. Assent, R. Krieger, T. Jansen, and T. Seidl, "Morpheus: interactive exploration of subspace clustering," *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2008, pp. 1089-1092.

[15] S.T. Teoh and K. Ma, "PaintingClass: interactive construction, visualization and exploration of decision trees," *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, Washington, D.C.: ACM, 2003, pp. 667-672.

[16] M. Ankerst, C. Elsen, M. Ester, and H. Kriegel, "Visual classification: an interactive approach to decision tree construction," *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 1999, pp. 392-396.

[17] Q. Cui and J. Yang, "Measuring Data Abstraction Quality in Multiresolution Visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, 2006, pp. 709-716.

[18] D. Yang, Z. Xie, E.A. Rundensteiner, and M.O. Ward, "Managing discoveries in the visual analytics process," *SIGKDD Explor. Newsl.*, vol. 9, 2007, pp. 22-29.

[19] G. Ellis and A. Dix, "Density control through random sampling: an architectural perspective," *Information Visualisation, IV 2002.*, 2002, pp. 82–90.

[20] E. Bertini and G. Santucci, "Give chance a chance: modeling density to enhance scatter plot quality through random data sampling," *Information Visualization*, vol. 5, 2006, pp. 95–110.

[21] R.A. Amar, J.T. Stasko, "Knowledge Precepts for Design and Evaluation of Information Visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 11, 2005, pp. 432-442.

[22] C. Plaisant, J. Fekete, and G. Grinstein, "Promoting Insight-Based Evaluation of Visualizations: From Contest to Benchmark Repository," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 14, 2008, pp. 120-134.

[23] J. Seo and B. Shneiderman, "A Rank-by-Feature Framework for Unsupervised Multidimensional Data Exploration Using Low Dimensional Projections," *Proceedings of the IEEE Symposium on Information Visualization*, IEEE Computer Society, 2004, pp. 65-72.

[24] P. Pirolli and S. Card, "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis," *Proceedings of International Conference on Intelligence Analysis*, 2005.

[25] J. Mackinlay, "Automating the design of graphical presentations of relational information," *ACM Transactions on Graphics*, vol. 5, 1986.

[26] D. Keim, "Visual Analytics: Combining Automated Discovery with Interactive Visualizations. (Invited Talk at VAKD'09 - http://www.hiit.fi/vakd09/keim.html)."