Surveying the complementary role 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 to integrate visualization and data mining techniques toward a mixed-initiative knowledge discovery taking 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 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 further uses this information to analyze the discovery process by comparing the analysis steps from the perspective of information visualization and data mining. The comparison permits to bring to light new perspectives on how mining and visualization can best employ human and machine skills.

Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: Graphical user interfaces (GUI). H.1.2 [**User/Machine Systems**]: Human information processing. H.2.8 [**Database applications**]: Data mining.

General Terms

Survey, Human Factors, Human-Machine Interaction.

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 [1][2] is an area of research where the integration of human and machine capabilities is advocated, beyond 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 an outgrowth of infovis and focuses on analytical reasoning facilitated by interactive visual interfaces [3]. Often, it is presented as being 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, it is not clear how this human-machine integration should happen. In our view, visual analytics should enable the collaboration between the natural abilities of humans and the powerfulness of data mining tools, thus combining in a synergetic way natural and artificial intelligences.

Despite the growing interests on this integration, however, 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 start shedding some light on this issue. 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 followed by a discussion of potential extensions we deem feasible and important to realize. The analysis is then followed 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 human and machine.

The paper is organized as follows. Section 2 introduces some terminology to clarify the meaning of some word 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. It describes the observed patterns and the potential enhancements we suggest. Section 5 dissects commonalities and differences between the analysis processes in data mining and visualization. Finally, Section 6 discusses the limitations of this work, and thus provides ideas for its future extension, and Section 7 closes the paper with conclusions.

2. TERMINOLOGY

The common goal of information visualization and data mining domains is to extract knowledge from raw data. Before going further in our inspection of this process, we thought useful to agree on the definitions of basic concepts that are commonly used in this context such as data, information, knowledge, model, pattern and hypothesis:

- *Data* refer to a collection of facts usually collected by observations, measures or experiments. Data consist of numbers, words, or images. It is generally called abstract data in infovis, since it refers to data that has no inherent spatial structure enabling further mapping to any geometry.

- A *model* in science is a physical, mathematical, or logical representation of a system of entities, phenomena, or processes. Basically a model is a simplified abstract view of the complex reality. Models are meant to augment and support humans reasoning, and further can be simulated, visualized and manipulated.

- A *pattern* is made of recurring events or objects that repeat in a predictable manner. The most basic patterns are based on repetition and periodicity.

- 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. The scientific method requires that one can test a scientific hypothesis. A hypothesis is never to be stated as a question, but always as a statement with an explanation following it.

- *Information*, in its earliest historical meaning, corresponds to the act of informing, or to the act of giving form or shape to the mind, according to the Oxford English Dictionary. Inform itself comes (via French) from the Latin verb "informare", to give form to, to form an idea of.

- *Knowledge* is the "justified true belief" according to Plato. According to the Oxford English Dictionary, knowledge can be defined as (i) expertise, and skills acquired by a person through experience or education; (ii) what is known in a particular field or in total; or (iii) awareness or familiarity gained by experience of a fact or situation.

In the context of *knowledge discovery*, we believe these concepts can be linked as follow: Data are the lowest level of abstraction; researchers often speak about *raw data* to emphasize this fact. 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 hidden nature of things or of perceiving in an intuitive manner. Patterns and models are not necessarily linked, even though some authors consider them as synonyms. One way to distinguish these two concepts 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 derived from models and patterns. 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 *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 library using keywords like: "data mining", "clustering", "classification", etc. Reversely, in data mining conferences we looked for the keywords like: "visualization", "interaction", etc. Manual skimming followed paper extraction. The final set of papers retained counts 55 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	VIS	V++	M++	VM
KDD	23	3	7	9	4
ICDM	16	2	5	5	4
INFOVIS	16	1	9	5	0

Table 1 - Distrubution 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

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:

- **Pure Visualization (VIS)** contains techniques based exclusively on visualization without any type of algorithmic support;
- **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 following we will not take into account the VIS category of pure 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 the technique becomes a "pure" visualization technique.

4.1.1 Observed enhancements with mining

As illustrated by black boxes on figure 1, the techniques collected in our literature review can be organized around three main patterns (Projection, Data Reduction, Pattern Disclosure) that represent different benefits brought by automatic computation to the information visualization process. Interestingly, as one can notice, the three patterns occur at the beginning of the knowledge discovery 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. The most traditional type of this method is Multidimensional Scaling (MDS), but in the literature it is possible to find many variations and alternatives. 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 on a more complex computation that takes into account all data dimensions and cases. Ward refers to this kind of placement techniques in [5] as "Derived Data Placement Strategies" in his glyph placement taxonomy.

- **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 the number of data dimensions by retaining subsets that carry the large majority of useful information contained in the data (and thus are most likely to show interesting patterns).

- **Pattern 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 useful patterns more easily emerge from the screen. Axes-reordering in parallel coordinates is one instance of such case [6]. Similarly, in visualizations where the degrees of freedom in visual configuration are limited, pattern detection algorithms can help make some visual patterns more prominent and thus readily visible. For instance, Vizster [7] organizes the nodes of a social network graph in automatically detected clusters enclosed within colored areas. Johansson et al. in [8] 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.
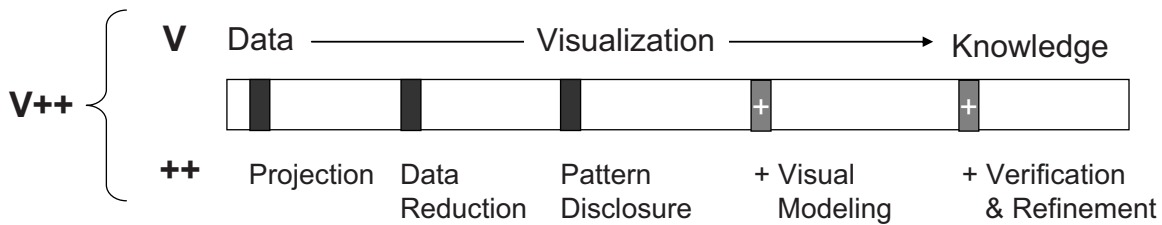

Figure 1 – Computationally enhanced Visualization (V++) benefit from mining techniques to improve information visualization standard process. Black boxes represent enhancements found in the literature survey; grey boxes (with "+") are extra benefits that could bring mining to visualization.

4.1.2 Other potential enhancements

All the automatic data analysis methods described above 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, as illustrated in figure 1 (grey boxes with "+"). Here we list some of the function we deem important:

- **Visual Model Building.** One limitation of current visualization systems is their inability to go beyond simple pattern detection and frame the problem around a scheme. 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* imply a functional model: any instantiation of the set of predictive variables returns a predicted 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 absent in visualization, nonetheless 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 [9] and the Perception Bases Classification (PBC) system [10] in which classification can be carried out interactively by means of purely visual systems.

- **Verification and Refinement.** One notable feature of automatic data mining methods 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 [11][12], 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 this issue [13][14] but none of them focuses on the use of sampling or data segmentation for verification purposes.

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 [15]. 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.

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.

4.2.1 Observed enhancements with visualization

As illustrated by black boxes on figure 2, 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. Interestingly, reversely to the previous category (V++), the two patterns occur at the end of the knowledge discovery process:

- **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
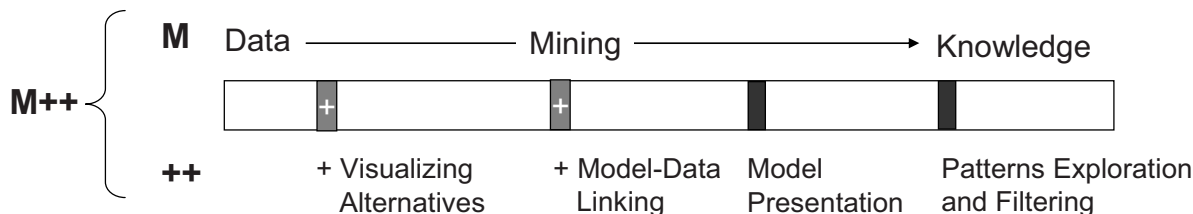

Figure 2 – Visually enhanced Mining (M++): benefits of visualization over data mining standard process. Black boxes represent potential enhancements found in the literature; grey boxes (with "+") are extra benefits that could bring visualization to mining.

support vector machines). Beyond 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 deeper understanding of the model and its underlying data.

- **Patterns Exploration and Filtering.** Some mining methods generate complex and numerous patterns which are difficult to summarize in a compact representation; notably 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 and overview of the distribution of these patterns and to make sense of their nature. Interactive filtering and direct manipulation tools have a prominent role in that finding the interesting pattern out of numerous uninteresting is the key goal.

4.2.2 Other potential enhancements

Visualization applied to data mining output, as shown above, 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 addressed so far, and enable users to intervene in early stages of the knowledge discovery process, as illustrated in figure 2 (grey boxes with "+"):

- **Visualizing Alternatives.** One of the characteristic features of data mining is the 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 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 blurred. Even more problematic, when alternative models are constructed it is extremely complicated to compare them in the space of a single user interface. Visualization in our opinion has the power to bridge this gap by: 1) providing means to more directly represent the connection between the parameters and the results; 2) allow for visualization structures that permit the comparison of alternative results. This last point is particularly 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 [16]. One system in our literature review partially supports this kind of comparison by generating different alternative results of a subspace clustering algorithm [17]. The user can see the results obtained through the variation of various parameters and choose the most interesting one among the set of available results.

- **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 often makes it difficult to interpret the observed relations in terms of the original data space. 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 seems to be the right tool to bridge this gap. One notable 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.

4.3 Integrated Visualization & Mining (VM)

This category combines visualization and mining approaches. None of them predominate 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. Following their description we speculate on a mixed-initiative approach to the KDD process.

4.3.1 Integration strategies

There are two extreme approached to integrate mining and visualization, as described below:

- **White-Box Integration.** In this kind of integration the human and the machine cooperate *during* the model building process in a way that intermediary steps in the algorithm can be visualized and decisions can be taken by the user on how to direct the model building process. This kind of systems is quite rare. There are examples of cooperative construction of classification trees, like the one presented in [18], 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. This kind of systems shows the highest degree of collaboration between the user and the machine and goes beyond the creation 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.

- **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 setting 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, could 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.

4.3.2 A mixed-initiative KDD process

Having analyzed a wide spectrum of integrations between automatic and interactive methods, 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 can cooperate on the same level.

Humans and machines are complementary, and visualization and data mining should make use of the specificities of each. Humans are intuitive and have good skills at interpretation according to the context and their domain knowledge. They are good at getting the "big picture" and at performing high level reasoning towards knowledge. Machine on the other side are fast and reliable at computing data, and they do not make errors.

In the early 90's already, Colgan & Spence et al. had already the vision to use visualization to enhance human-machine collaboration in electronic circuit design through the cockpit of their Coco system. Their approach highlighted the need for an effective interface to blend the complementary capabilities of the human designer and computer algorithms [22, 23]. More recently, Pu & Lalanne proposed a mixed-initiative system to support problem solving via algorithm visualization and visual trade-off analysis [20, 21]. Through visual interaction, the Comind system enables designers to select and control the solving algorithm they want to use, i.e. they can visualize it while it is processing the data, stop it at anytime and modify the problem definition or select another mining or solving algorithm. Finally they can select the visualization techniques they want to view the results, while still being able to tune parameters. In the context of sequential pattern detection for text mining, [19] proposes to combine computational and statistical efforts through data mining with the human participation through visualization for the ultimate goal of knowledge discovery. In their application, visualization helps humans quickly obtain an overall structural view of patterns and complementary, data mining provides accurate support information for all patterns.

Table 2 summarizes the major complementary strengths of human and machine in the knowledge discovery process, derived from our literature review.

Human	Machine
Select strategies	Project & Reduce data
Observe, derive knowledge	Select optimal solution, best configuration
Interpretation, explanation	Build models
Measure interestingness	Extract patterns, models
Generating hypothesis	Verification

Table 2 – Complementary strengths of human and machine in the knowledge discovery process.

Figure 3 is the result of the benefits brought by visualization and mining independently to the knowledge discovery process as described in section 4.1 and 4.2 respectively. It is inline with the complementary strengths brought by humans through visualization and by machines through data mining. For example, while humans are good at choosing modeling strategies through visualization, the machine is good at computing large amount of data for projecting and reducing data. Further, while machines can disclose and highlight all the patterns found automatically over the data, human can explore them and keep only the most interesting ones, according to their knowledge of the data set and its associated domain. Later on, 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 [18], where a decision tree can be constructed by alternating steps of human-based decisions and machine-based algorithmic steps.
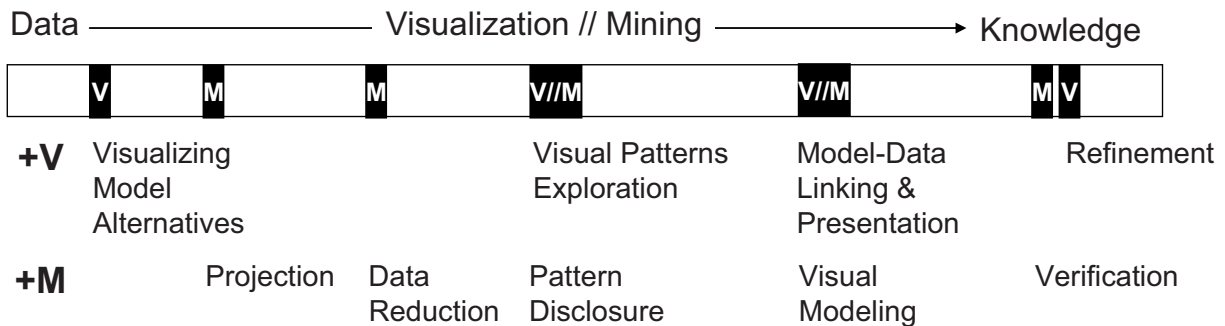
Figure 3 – Integrated visualization and Mining (VM): towards a white box for the full KDD process with benefits coming from visualization (+V) and benefits from mining (+M).

5. ANALYZING THE ANALYSIS PROCESS

Both visualization and data mining are alternative methods to transform data into knowledge. Having said that, a legitimate question remains: are they just different recipe that work in the same manner or do they differ in any substantial manner? We believe that posing this question is becoming of increasing importance as we attempt to get the most out of the two and create successful integrations like the one advocated in Visual Analytics.

Here we provide reflections on this subject, based on an initial schematization of the analysis process in data mining and visualization, highlighting notable differences and commonalities between them.
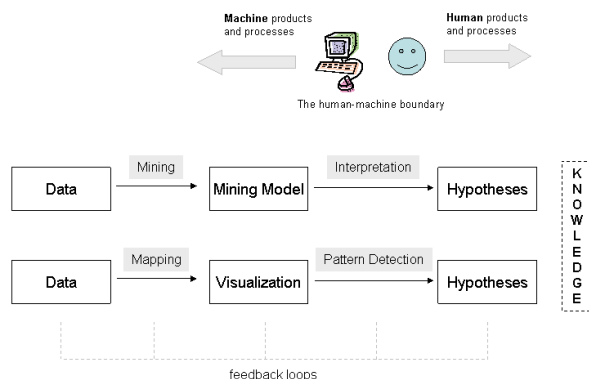

Figure 4 - Comparison between mining and visualization analytics processes.

5.1 Processes versus products

Looking at Figure 4 we can see that both in visualization and mining we have products (boxes) and processes (arrows). What is interesting to note, at least from the terminological point of view, is that *visualization* and *data mining* are not on the same level. More precisely, the word "visualization" is often intended as the *product* of the visual mapping between data and a visual representation; the word "mining", on the other hand, commonly refers to the *process* that transforms data into a data mining model. This distinction is important because in Visual Data Mining and Visual Analytics often mining and visualization are considered as alternatives. Even more important is to acknowledge the fact that in data mining there are necessarily always some tasks performed by the human and, likewise, in information visualization there are always some tasks performed by the machine. The machine, in particular, is responsible of the *mining process*, in data mining, and of the *visual mapping process*, in visualization. Moreover, the mining process produces a *mining model*, whereas the visual mapping process produces a *visualization*.

If we adopt this perspective it is easy to see for instance how in visual mapping and mining process similar human tasks are involved, like the definition of an appropriate schema (visual or functional) that fits the user's mental model and goal. Similarly, we realize that in terms of perceptive and cognitive processes it is the comparison of the activities that go from visualization to hypothesis generation, in visualization, and from mining model to hypothesis generation, in mining that matters. We believe that a deeper analysis and comparison of what happens at this stage, where the human interfaces with the machine, might lead to relevant advancements in Visual Analytics.

5.2 Mental models and problem instantiation

Again, comparing the two processes in Figure 4, it is interesting to note a key difference between them. In visualization the formation of a mental model and its formalization happen "in sequence" when the mapping has already been performed and the data is already visualized. In other terms, the visualization by itself is a vehicle to aid the formation of a mental schema. In data mining instead the human has to first *mentally* formulate a mental schema in a way that it can fit with one of the existing input-output mappings provided by data mining.

A clarifying example comes from the comparison of how knowledge building happens on Parallel Coordinates visualization or a Decision Tree algorithm. In the first case, the user most probably approaches the problem with a limited formalization of the problem space and an opportunistic approach. Usually he or she just wants to look at the data and see what's there. Moreover, the kind of extracted patterns can cover a quite broad range of data models, e.g., correlations among two or more dimensions, groupings (clusters), outliers, etc. In the case of decision trees, the user has to first formulate the problem in terms of a definite mental schema that matches the particular input-output mapping enforces by the technique. Specifically, the data will be transformed in a series of IF-THEN rules that segment the input space in groups characterized by their relations. For any additional data record, once the model is built, the model will provide a specific output (label). It is worth to note in this example that some of the conclusions to which the user might end up in one case might easily overlap with those extracted from the other. The question of how these processes compare, when and how it is more preferable to use one or another, or if a synergy between the two can be found is in our opinion one of the central issues to study in Visual Analytics.

5.3 The Human-Machine Interface

Another important aspect illustrated in figure 4 is that in both processes there is a stage in which necessarily the human has to acquire some information from the machine, that is, what we called the *human boundary*.

In traditional data mining, systems are not without an interface, they just provide simple and minimalistic interfaces like results organized in tabular data. Visualization systems on the other hand provide visually rich and highly interactive tools for data exploration.

More importantly, in data visualization the interface has the primary goal to let the user *detect* and correctly extract relevant patterns from the screen. In data mining the interface has the primary goal to let the user *understand* the model produced by the machine and its relation to data. From the visualization design point of view it is important to recognize this difference and acknowledge that not necessarily what we have learned from data visualization is enough to build effective model visualizations.

Model visualization seems to be a more complex task, where we are confronted with novel design challenges like: finding effective metaphors to represent the model, finding ways to represent the

model in relation to the data and vice versa, and finding convenient interaction methods to manipulate the model. Further research is still needed to advance towards this direction.

5.3.1 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 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 this state and devise alternatives that fuel the discovery and learning process.

On a higher level this is also supported by the Sensemaking Theory that describes how people make sense of information. As Pirolli and Card note in [19], the process revolves around "one set of activities that cycle around finding information and another that cycles around making sense of the information".
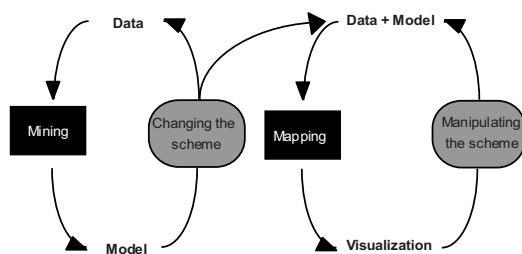

F**igure 5 – The feedback loop in Knowledge Discovery. The grey boxes represents the two major stages at which humans can intervene**.

5.3.2 User intervention levels

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

- **Changing the scheme.** 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 an example, 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; especially those found in information visualization. One of the biggest challenges is yet 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 [20]), 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 scheme.** Another option the user has to create alternatives is to change parameters within the context of a given scheme. In visualization this comprises 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 4.2.2 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.

6. 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 complete. We decided to use a number of papers that could be analyzed in a relative short time (by the two authors) and at the same time capture most of the relevant trends.

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 general trends out of it. We postpone this task to a later version of our work, where the number and kind of collected papers will provide us with a more solid base on which to draw relevant statistics.

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. Specifically, we want to extend the literature, further categorize the techniques, and draw some general statistics on research trends that could help suggesting additional future research directions.

7. 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 techniques. In particular, this preliminary 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 scheme respectively.

8. 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]   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.

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

[8]   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.

[9]   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.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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