Machine Learning Methods in Visualisation for Big Data (2018)
D. Archambault, I. Nabney, and J. Peltonen (Editors)

Panning for Insight: Amplifying Insight through Tight Integration of Machine Learning, Data Mining, and Visualization

Benjamin Karer, Inga Scheler and Hans Hagen

TU Kaiserslautern, Germany

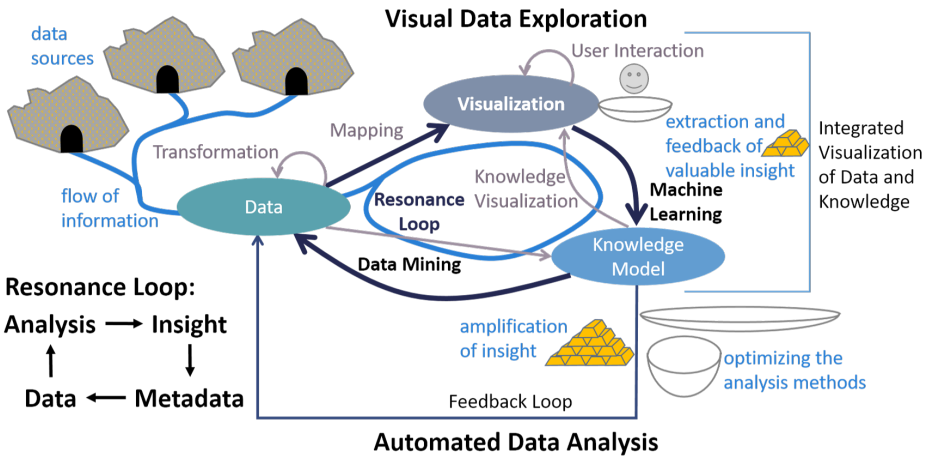

Figure 1: *The analogy of panning for gold in a river illustrates how the proposed analysis workflow optimizes the extraction of insight. A feedback loop makes insights obtained in previous analysis steps directly available to automated and human analysis.*

Abstract

With the rapid progress made in Data Mining, Visualization, and Machine Learning during the last years, combinations of these methods have gained increasing interest. This paper summarizes ideas behind ongoing work on combining methods of these three domains into an insight-driven interactive data analysis workflow. Based on their interpretation of data visualizations, users generate metadata to be fed back into the analysis. The resulting resonance effect improves the performance of subsequent analysis. The paper outlines the ideas behind the workflow, indicates the benefits and discusses how to avoid potential pitfalls.

CCS Concepts

•***Human-centered computing*** → ***Visual analytics;*** *Visualization theory, concepts and paradigms;*

1. Introduction

Driven by advances in data mining, machine learning, and visualization, decentralized data collection, aggregation, integration, and analysis became almost ubiquitous. The advantage of making the results of automated data analysis accessible for human interpretation is well documented by the success of visual analytics. Yet, the obtained insights commonly remain entirely with the human analyst. This paper aims at exploiting this rich source of information by an improved visual analytics pipeline.

The paper's contribution is an iterative workflow, merging the data model and a model of the knowledge aggregated during analysis to perform analysis directly on this model rather than on the data. This way, insights can not only be obtained from analyzing the data directly but also from the evaluation of previously obtained results, which by the model integration are available for automatic analysis. Machine learning is applied to make the analysis feasible for large data sets by quickly rolling out insights found locally for some structure to similar structures in the data.

DOI: 10.2312/mlvis.20181130

2. Searching for Insight and Panning for Gold

If data is too large to be processed at once or analysis is to be performed in an online setup constantly generating new data, the analysis is often limited to comparably small subsets of data being streamed through the system. Having to decide on which subset of the data is to be evaluated, one could claim that insight is only worth as much as the added value it generates. For the search of valuable information in streaming data, the metaphor of searching for a needle in a haystack often applied in big data contexts in some sense translates to panning for gold in a river.

Data Mining, Visualization, and Machine Learning each have their own approaches to finding information. Using explorative visualization, the user would attempt to find a spot along the river where the yield of panning for gold is maximized. This could very well require to explore the whole river. Data mining would try to analyze and cluster particle patterns in the stream. The interpretation where to find the gold and how to extract it from the stream is left to the user. Sophisticated machine learning algorithms would find an efficient strategy to extract large amounts of gold – if they were trained properly. If the training data is not of sufficient quality, the algorithm might as well just extract tons of sand.

A combination of the three approaches could for example proceed as follows: The data mining's clustering is interpreted by the user by means of visual analytics. The most promising streams are bundled by canals and led into a cycle to increase the potential yield even further. The gold to be extracted has a specific shape and floating behavior. Panning for some gold and labeling particles accordingly yields training data for machine learning. In some sense, learning to keep the gold and let the other particles pass in an optimal manner can be thought of as optimizing the pan. The result is an optimized gold extraction procedure to be applied to the water stream. Feeding the gold obtained from panning back into the system thus results in an accumulation of more gold and better yield. A larger gold yield means an improved return on invest. For data analysis, this means an improved efficiency of insight obtainment.

3. Related Work

The integration of automated data analysis and visualization is the fundamental idea behind visual analytics. The classical visual analytics pipeline as proposed by Daniel Keim [KAF*08] is well reflected in existing systems for interactive data analysis. A survey conducted in 2016 reveals that most visual analytics pipelines follow this principal scheme and specialize certain aspects [WZM*16]. This paper proposes an alteration of the pipeline, merging the knowledge and data models and thereby augmenting the data being analyzed by the insights.

There is a variety of tools offering to combine Data Mining and Visualization. Some of them also include Machine Learning algorithms. KNIME and Orange are only two of the more well-known examples [BCD*07, DCE*13]. These tools typically offer a graphical interface for the specification of data processing pipelines and visualization to study the results of pipeline executions. However, they do not feature the direct reintegration of obtained insights into the data proposed in this paper. A recent survey reviewed 19 open source tools for data mining with respect to their quality and their features [ALVV17]. While most of the tools provide a visualization of the resulting model, less than half of them offer to visualize the data. Only about half of the tools (10/19) allow saving and reloading the results and only five can export the obtained results to common exchange formats like XML. Since the workflows in these tools are typically implemented as unidirectional linear or tree-like structures, saving and reusing obtained models is a necessity to implement an iterative approach like the one proposed in this note. Most of the reviewed tools are focused on the construction of data processing pipelines. This work, instead, focuses on the data itself, especially on the metainformation obtained by the user who interprets the visualization. Putting the focus on reintegrating obtained insight into the data increases the resource requirements. For this kind of scaling problems, Starič et al. recommend to work with light-weight visualizations supporting the parallel and asynchronous execution of algorithms [SDZ15].

The most relevant related work to the proposed workflow is the human-centered Machine Learning framework proposed by Sacha et al. [SSZ*17]. Similarly to the approach proposed here, an iterative workflow based on Keim's Visual Analytics model is discussed, where the user applies domain knowledge to steer Machine Learning algorithms to support the analysis process. The paper also provides a good overview over existing approaches implementing parts of such a pipeline along with an in-depth discussion of tasks and analysis steps to be performed in such a setup. In their discussion, Sacha et al. focus on interaction for model building and paramater refinement to improve the performance of Machine Learning algorithms by leveraging the user's domain knowledge. In contrast, the approach proposed here is focused on restructuring and augmenting the data than on the refinement of parameters and model definitions. Altering the analysis pipeline as illustrated in Figure 2 directly integrates the analyst's mental model with the available data. Taking into account insights obtained from previous analysis steps effectively extends the capabilities of the framework proposed by Sacha et al.

4. A Resonance Loop Amplifying Insight

In the classical model of visual analytics proposed by Daniel Keim (cf. Figure 2), the user applies interactive visualization and data mining to build, verify and refine a data model. The additional information offered by the model generates an added value for the interpretation of the visualization providing insight into possibly hidden relationships and dependencies in the data. Like in the thought experiment outlined in Section 2, this process can be seen as a flow of data (particles) along a stream where different means of analysis (the pans) are applied to extract valuable insight (gold). Although Keim's model includes the notion of a feedback loop from knowledge to data, this loop is only of conceptual nature and indicates the idea that users may choose to concentrate on different data based on the knowledge obtained from previous analysis [KMS*08]. The obtained insight remains out of system, rendering the extraction of knowledge essentially unidirectional.

It is not uncommon that the information cannot be read off directly but has to be inferred by reasoning about multiple data elements. Being aware of this problem, visual analytics applies Data Mining to obtain data models in which the information can be

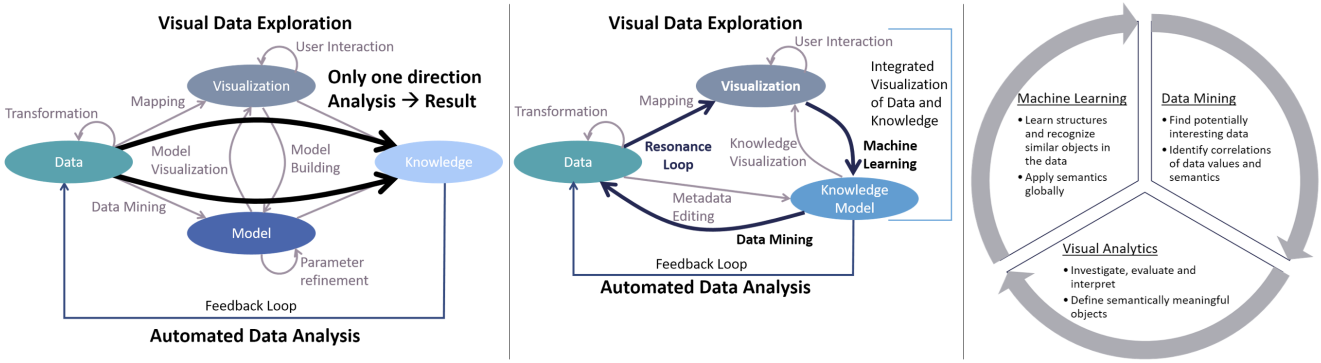

Figure 2: *Daniel Keim's model of visual analytics (left), the modification proposed in this paper (center), and the proposed analysis workflow (right). Most existing visual analytics applications follow the classical unidirectional flow of knowledge from the visualization and a data model to the viewer. Merging the knowledge and the data model implements an augmentation of the data being analyzed by the insights obtained from prior analysis steps. Leveraging this augmentation mechanism, an iterative workflow combining methods from Data Mining, Visualization, and Machine Learning, implements a resonance loop fostering the obtainment of new insights from previously obtained results.*

found more easily than from studying only the raw data. Finding the information might induce a new analysis question. An unsuccessful user instead could apply interaction to edit parameters steering the preprocessing, or decide to investigate different portions of the data or an entirely different data set. This is the feedback loop in Keim's model. Patterns, redundancies, or other interesting observations might not only be hidden in the data's values but also in the interpretation. Sometimes, the information to be found within the data but is hard to detect. This is the case, for instance, if the information is to be derived from transformed data, for example from the derivatives of a scalar field rather than from the field itself. If the derivatives are not part of the data, the feedback loop proposed in this paper instead allows to evaluate the derivatives in local neighborhoods and to label the resulting new data accordingly. Investigating the derivatives and identifying the interesting information, the user could now select and label the respective derivative data. Machine Learning can be applied to roll out these findings to the rest of the derivative data and the parameters steering this process can be optimized by a workflow similar to the one proposed by Sacha et al. [SSZ*17]. Each derivative value can be mapped back to the original data points which can now be evaluated with respect to the insights found in the derivatives.Rather than only considering different data, the user thus concentrates on different information associated with the data which is made possible by including the knowledge obtained about the derivatives into the data and aligning it with the original data.

The visual analytics pipeline is thereby transformed into a resonance loop, amplifying the generation of new insight. An illustration of this model compared to Keim's model is shown in Figure 2. In the analogy of panning for gold, the feedback loop maps to the application of data mining to finding promising data sources (rivers with high yield) and integrating them properly in a preprocessing step (channeling the flow). The user derives insights from interpreting visualization (the gold obtained from manual panning) and feeds the results back into the system as metadata. Machine learning is applied to iteratively refine the data and knowledge model (optimizing the pans). With the assistance of automated procedures to mine and analyze previously obtained knowledge and apply it to the data, new insight can be derived from previous results (amplifying insight). Figure 1 illustrates the thought experiment overlaid with the proposed analysis workflow.

Optimal results require a tight integration between the three domains in a workflow leveraging each field's specific strengths and alleviate each other's weaknesses. The purpose of Data Mining is the detection of previously unknown patterns in the data. Their interpretation is left to a human analyst. Visualization is an interface for humans to make sense of data. Yet, finding and interpreting structure requires a skilled user and often also considerable amount of time. In direct comparison to Data Mining, the focus of Machine Learning is more on the identification of patterns already known. Its results, however, rely heavily on the proper choice of training data. Assigning roles to the three domains according to these strengths and weaknesses implements the workflow illustrated to the right of Figure 2. While on the global scale the proposed workflow implements a loop of applications of Data Mining, Visualization, and Machine Learning, each executed procedure is based on an individual linear transformation pipeline. There is a variety of open source tools available for the creation of such pipelines [ALVV17].

The metadata to be edited can take multiple forms. Perhaps the simplest method to map analysis results back to the data is the assignment of labels. There are no strict restrictions to the data's shape other than that it needs to be compatible with the applied Data Mining and Machine Learning algorithms. Since the metadata is meant to formalize insights found during analysis and these insights will typically be of a qualitative, descriptive nature, it makes sense to apply a data structure explicitly mapping sets of data items to semantic information. The transformations applied during data processing and interaction organize this structure in a graph allowing to navigate the analysis results obtained so far. If in such a setup the applicable Data Mining and Machine Learning procedures offered by the system are known for every set of data items, pipelines processing the data to serve complex information queries

can be generated automatically [KSH18]. In its most simple form, the metainformation is simply a set of labels applied to the respective set of data items. However, more complex structures like a semantic web or other kind of ontology defined on top of the data are feasible and allow more sophisticated analysis and inference structures operating directly on the knowledge model.

To make the assignment of labels or other metainformation feasible for large data sets, a semiautomatic distribution of metainformation can be achieved by searching for data patterns to be labelled rather than for individual data items. Depending on the shape of the data patterns, suitable Machine Learning algorithms can be trained using the labeled data to roll out the labels to corresponding structures in the remaining data. As an example, consider a point cloud obtained from scanning, for example, an asteroid's surface. Due to measurement errors, there is some noise in the data and the surface is not smooth. While the analyst would be interested in studying craters, the measurement errors induce false local critical values. Simply smoothing or averaging the surface could, however, result in the loss of important detail. Data mining can be applied to categorize local neighborhoods of points with respect to the points' position relative to an averaging surface. The clusters will reveal bumps, dents, ridges, and other structures. For the analysis of craters, too small neighborhoods result in a large number of erroneously found crater-structures whereas the cluster criterion does not yield reliable results for too large neighborhoods. A simple application of machine learning would be to find craters by searching for the largest structures whose similarity to a local bump or dent does not fall below a certain threshold.

5. Example Use Case

Irregular influences on air-traffic patterns like thunderstorms do not follow spatial patterns. Their influence on air-traffic routes can thus not be accurately predicted based on historical data. Nevertheless, historical data can be considered to identify possible evasion routes. The following discussion shows how the proposed workflow could be applied to solve this problem by mapping each analysis steps to the domains of Data Mining (DM), Visualization and Visual Analytics (VA), and Machine Learning (ML).

If a storm warning is announced, historical data is mined for past storms in the same region (pattern recognition, DM, channeling streams). The user assigns grades to the trajectories of representative planes evading the storm to assign them to equivalence classes reflecting their quality (find and evaluate structures, VA, manual panning for gold). These grades are now rolled out to the other evasion routes by a classification algorithm (classification, ML, optimizing pans). Quality measures determine how well each path fits into its class (cluster quality assessment, DM, determine yield quality). Where necessary, the identified classes are subdivided into two or more subclasses by assigning proper labels (evaluate quality and detect subclasses, VA, increase the gold yield). These adjustments to the classifier's definitions improve the results during reclassification (reinforcement learning, ML, optimizing the pans). The controller identifies the best-graded routes for every relevant direction and reevaluates their embedding into the actual surveillance data (VA, panning for gold). The planes can then be assigned to the evasion routes according to the classifier trained before (ML, increase yield). Storing the routes for future reference, candidates can be obtained directly from the collection rather than having to be extracted them from historical data (amplify insight).

6. Avoiding Credibility and Reliability Issues

Feeding back the results of visual analytics to into the data to make it accessible for Machine Learning and Data Mining enables the derivation of new insight from previously obtained results. With the benefits, there also come pitfalls and risks. In visual analytics workflows, uncertainty usually only propagates between the data and the obtained model from the data and the model to the visualization [BHJ*14]. Feeding back analysis results into the data and the model introduces two additional types uncertainty: a quantitative uncertainty in the classification obtained from machine learning and a qualitative uncertainty regarding the credibility and reliability of the results obtained from human data analysis.

Other than the human user, the computer does not reflect on the data it receives as input. Thus, errors in the analysis will not be detected by the computer and propagate through further computation. When attempting to roll out analysis results to the whole data set, misclassification errors can be corrected by refining the classification schemes. Still, there is a risk of an "analyst-induced oscillation" where continued optimization attempts eventually result in an overfitting detrimental to the classifier's performance.

To assess the credibility of metadata defined in previous analysis, provenance information must be stored along with the metadata. Without such information, errors made in previous steps or assumptions inapplicable to the current investigation might yield false analysis results. Note that, being part of the metainformation added to the original data, the provenance information can be accessed and processed like any other data.

To test the model's reliability, it can be tested against the addition of new (artificial) data and against assertions. The metadata and definitions together define a model for the observation. If the model is accurate, it should predict the metadata of newly added data points correctly by applying the definitions obtained from previous analysis. Assertion checks can be performed by specifying a condition that has to hold under the model. This assertion is then evaluated on each relevant data item generating a label with the evaluation's result. The labels can then be used for further analysis to check whether the assertion holds on the correct data elements.

7. Conclusion

This paper proposes an extension of the visual analytics pipeline achieving a tight integration of Data Mining, Visualization, and Machine Learning. Where in the original workflow the information obtained from automatic analysis is available for the human analyst but the insights obtained are not meant to be processed by the computer, the new workflow applies Machine Learning to close this gap. The resulting iterative workflow leverages the three domains' respective specific strengths to foster the obtainment of new insights from previous results. Implementations of this workflow can be expected to increase the efficiency of data analysis, yielding more sophisticated insight in less time.

References

[ALVV17] ALTALHI A. H., LUNA J. M., VALLEJO M. A., VENTURA S.: Evaluation and comparison of open source software suites for data mining and knowledge discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 7*, 3 (6 2017). 2, 3

[BCD*07] BERTHOLD M. R., CEBRON N., DILL F., GABRIEL T. R., KÖTTER T., MEINL T., OHL P., SIEB C., THIEL K., WISWEDEL B.: KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)* (2007), Springer. 2

[BHJ*14] BONNEAU G.-P., HEGE H.-C., JOHNSON C. R., OLIVEIRA M. M., POTTER K., RHEINGANS P., SCHULTZ T.: Overview and state-of-the-art of uncertainty visualization. In *Scientific Visualization*. Springer, 2014, pp. 3–27. 4

[DCE*13] DEMŠAR J., CURK T., ERJAVEC A., ČRT GORUP, HOČEVAR T., MILUTINOVIČ M., MOŽINA M., POLAJNAR M., TOPLAK M., STARIČ A., ŠTAJDOHAR M., UMEK L., ŽAGAR L., ŽBONTAR J., ŽITNIK M., ZUPAN B.: Orange: Data mining toolbox in python. *Journal of Machine Learning Research 14* (2013), 2349–2353. 2

[KAF*08] KEIM D., ANDRIENKO G., FEKETE J.-D., GÖRG C., KOHLHAMMER J., MELANÇON G.: Information visualization. Springer-Verlag, Berlin, Heidelberg, 2008, ch. Visual Analytics: Definition, Process, and Challenges, pp. 154–175. 2

[KMS*08] KEIM D. A., MANSMANN F., SCHNEIDEWIND J., THOMAS J., ZIEGLER H.: Visual data mining. Springer-Verlag, Berlin, Heidelberg, 2008, ch. Visual Analytics: Scope and Challenges, pp. 76–90. 2

[KSH18] KARER B., SCHELER I., HAGEN H.: A step towards automatic visual analytics pipeline generation. In *Electronic Imaging 2018: Visualization and Data Analysis* (2018). 4

[SDZ15] STARIČ A., DEMŠAR J., ZUPAN B.: Concurrent software architectures for exploratory data analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 5*, 4 (7 2015), 165–180. 2

[SSZ*17] SACHA D., SEDLMAIR M., ZHANG L., LEE J. A., PELTONEN J., WEISKOPF D., NORTH S. C., KEIM D. A.: What you see is what you can change : Human-centered machine learning by interactive visualization. *Neurocomputing 268* (2017), 164–175. 2, 3

[WZM*16] WANG X.-M., ZHANG T.-Y., MA Y.-X., XIA J., CHEN W.: A survey of visual analytic pipelines. *Journal of Computer Science and Technology 31*, 4 (Jul 2016), 787–804. 2