Towards Visual Data Science - An Exploration

Marina Tropmann-Frick and Jakob Smedegaard Andersen(✉)

Department of Computer Science, Hamburg University of Applied Sciences, Hamburg, Germany
{marina.tropmann-frick, jakob.andersen}@haw-hamburg.de

Abstract. Visual perception is one of the most essential abilities for humans. This ability allows us to discover the world around us and to understand interdependencies with regard to both global context and particular concrete problem statement. We present in this paper an exploration of visualization and visual analytics approaches. Thereby, we focus on the field of data science with data-centric analytic methods and applications. Data science is closing the gap between visualization techniques, traditional hypothesis-driven methods and processing of mostly huge, heterogeneous and noisy data. A combination of smart visualization, advanced analytical methods and additional (domain) knowledge, mostly provided by humans, makes it possible to gain insights and discover new opportunities for problem solving.

Keywords: Data science · Visualization · Visual analytics · Visual interaction

1 Introduction

Data science is a relatively new rapidly developing field dealing mostly with big data. The goal of data science applications is to analyse and to extract knowledge from data, which can be used for simulation, decision making, process optimization, innovation support, etc. The process of knowledge discovery comprises data preparation techniques, statistical modeling, computational methods, visual analysis, and domain-driven problem solving. Data science relies on methodology developed in such fields as statistics, data mining, machine learning, big data, or human–computer interaction. Visualization and visual analytics build an essential part of data science.

Although both domains are dealing with visual representation, their scope and impact are different. Visualization provide techniques for presentation of data or relationships for the purpose of explanation, interpretation, communication etc. Visual analytics encompasses a process of knowledge discovery by supporting the analyst to discover patterns in data, building formal models that can be processed by machines, and developing new hypothesis. Visual analytics strongly relies on effective interaction of a human and a machine. The *Human in the Loop (HiL)* concept is characteristically for the continuous support of machine processing by human feedback. In the context of Visual Analytics, *HiL* stands for providing continuous feedback, correcting algorithmic approaches and selecting appropriate techniques during the analytical process.

We present in this work a survey on visual analytics in data science. The next section describes the main visualization aspects. Then we discuss the domain of visual

T. Ahram et al. (Eds.): IHIET 2019, AISC 1018, pp. 371–377, 2020.
https://doi.org/10.1007/978-3-030-25629-6_57

analytics and take a closer look on the *HiL* concept as well as present a comparison of visual analytics tools developing by experts and described in scientific publications in computer science journals.

2 Visualization and Visual Analytics

The design space of visualization is enormous ranging from aesthetic aspects to structural organization that affect perception. Many proven techniques have been developed to present information. These include various types of diagrams for graphical representation of proportions or numerical values as well as structures for mapping relationships and set affiliations. Visual processing pursues several different objectives such as the recognition of contexts, patterns and trends, finding outliers and clusters, classification or examining structures. Data objects can be integrated in a visualization in different ways, such as points, lines or areas as a representation object over the visual properties such as colour, shape, size and position, and the layout of the presentation.

Visualization should always be done very carefully. It is easy to misinterpret or to get misleaded by charts and graphs. Misleading data visualizations are often designed by violating standard practices. People are used to the fact that pie charts represent parts of a whole or that timelines progress from left to right. When those rules get violated, we would very likely misinterpret such a visualization. Nonetheless, a non-standard technique can make a visual representation more expressive and give it an outstanding position. Such visualizations draw our attention and make us aware of the subject representing.

Visual analytics is an interdisciplinary approach to support exploratory knowledge discovery especially regarding large and complex data sets [1–3]. The subject of visual analytics is the effective obtaining, extending and generation of knowledge. The central concept of visual analytics is to combine automated analysis techniques with interactive visualization in order to improve the overall analysis [2]. Through combining these fields, more challenging problems can be solved, which cannot be addressed by using only pure analytical or visual approaches. On the one hand, the data is often too large for pure visual approaches and therefore requires an automated pre-processing step. On the other hand, many problems are too exploratory in nature for purely automated analysis, which demands for the integration of human cognition.

Visual analytics is primarily concerned with the needs of exploratory data analysis [1]. Problems, which are solved by visual analytics, are initially unknown or vaguely formulated and presume a specific background knowledge, which the machine does not possess [4]. Through the integration of the user in an early stage of the analysis process, the user gets into the position to interactively learn the data and thus to specify and set up hypotheses. The exploratory orientation of visual analytics assists the analysts in unveiling unexpected connections and phenomena [3]. The analyst can follow various intentions in different sections of the analysis like overlooking the data, the exploratory search for new insights or the testing of hypotheses. In addition, the analysis process ranges from high level analytical tasks which strongly depend on background knowledge and expertise of the user to low level activities like the selection of the underlying data sets.

2.1 Human in the Loop

One of the main concepts regarding interactive analysis is the *HiL* concept. *HiL* emerged from the observation that some machine approaches require analytical judgement, others can be significantly improved or accelerated by interacting with humans. Characteristically for *HiL* is the continuous support of machine processing by human feedback. In the context of visual analytics, this concept occurs in terms of providing continuous feedback and correcting algorithmic approaches within the analysis. There is no general accepted definition of *HiL* and it remains open, which loop is finally meant. The optimization of learning behaviour in machine learning is a field of application for *HiL* in visual analytics, which is often used. The user is hereby directly integrated in the train-tune-phase as the user implicitly or explicitly alters the parametrization. For instance, the user extends the underlying training data, corrects them or might provide additional information.

The combination of analytical reasoning and computational models can result in usability problems. Problems arise for instance if interactions are not intuitive. Often users have to express their expectations through a variety of parameters and configurations. Users are confronted with the issue to communicate their knowledge and expertise with the machine, which affects the overall process negatively. Therefore, Endert et al. argues for a shift towards a *Human is the Loop* perspective to focus more on a seamless integration of interactions [5]. There is a demand for orientating interactions more on analytical cognitive processes in order to reduce cognitive burdens.

Nonetheless, human reasoning, decision making and knowledge generation processes are essential for the effectiveness of *HiL* and should have a central role in the process. We consider the loop as part of the computational sense making by means of adjustable environments. In our opinion *HiL* describes a machine environment that is managed by human knowledge in order to conduct continuous analytical discourses.

3 Comparison

As we mentioned above, visualization and visual analytics represent two different approaches to gain insights from data. Visualization is concerned with the depiction of data to assist the perceiving of patterns, structures and coherences. The objective hereby is to gain insights in order to comprehend data, make decisions and to build trust in the underlying data. Data as well as models are depicted through visual artefacts or diagrams. In order to visualize data effectively, it has to be in a structured form. In this context pre-processing steps like data cleaning and wrangling are applied on the raw data. Visualizations are not just static, but also provide interactions to change views.

The main focus of visual analytics is on the solving of problems. Visual analytics uses the techniques and methods provided by visualization to close the gap between algorithmic processes and human factors. Visualization is primarily concerned with the

selection of optimal visual forms and interactions for a given problem, whereas visual analytics deals with the integration of analytical approaches into the entire knowledge generation process. The challenges of visual analytics are to find the best analytical approaches for a given problem, automate them as much as possible and finally provide an integrated tool, which considers human factors [2].

Since finding the right models and parameters for a given problem can be a difficult task, this task can also be shared with the user. In comparison to visualization, visual analytics includes a structured reasoning process to gain knowledge out of data [6]. The user plays an active role by steering the underlying processes and models via interactive interfaces. Visualization is used to continuously clarify the state of the analysis, to visualize data and to communicate new knowledge mutually. The combination of data-centric methods with user-centric interactions through visual interfaces is an object of visual analytics. Whereas visual representations without reference to analytical approaches are the subject of visualization.

Table 1. Comparison of visual analytics tools from (top down) [7–26]

	Why				Who			What						How		
Visual Analytics Tools	Exploring Data	Interpretability & Explainability	Debugging & Improving Models	Comparing & Selecting Models	Model Developers & Builders	Model User	Non-Experts	Relationships	Similarities	Time	Subsets	Aggregated Information	Distribution	Points	Lines	Areas
Kwon et al., 2018	x	x		x		x			x			x	x	x	x	x
Stasko et al., 2007	x	x					x	x			x	x		x	x	x
Cavallo and Demiralp, 2019	x	x	x	x	x	x			x		x	x	x	x	x	x
Park et al., 2018	x	x				x			x	x	x	x		x		x
Gotz and Stavropoulos, 2014	x	x				x	x	x	x	x	x	x	x	x	x	x
Höferlin et al., 2012		x	x	x	x							x		x		x
Jeong et al., 2009	x	x	x		x	x			x		x	x	x	x	x	
Brown et al., 2012	x	x	x		x	x			x			x	x	x		x
Alsakran et al., 2011	x	x	x				x		x	x	x		x	x		
Soo Yi et al., 2005	x						x		x		x			x		
Krause et al., 2014	x	x	x	x	x	x					x	x	x			x
Strobelt et al., 2018	x	x				x			x	x	x	x			x	x
Bögl et al., 2013		x		x		x				x	x	x	x	x	x	x
Krause et al., 2016	x	x					x	x			x	x	x	x	x	x
Chandrasegaran et al., 2017	x	x					x			x	x		x		x	x
Du et al., 2016	x	x					x		x	x	x	x	x	x	x	x
Kim et al., 2017	x						x		x		x	x		x		
Kwon et al., 2017	x					x			x		x	x			x	x
Pezzotti et al., 2018		x	x		x					x		x		x	x	x
Bailey et al., 2018	x						x	x		x	x	x			x	x

Visual Analytics focuses on exploratory analysis and is not limited to visualization and automated analysis. It also includes the entire infrastructure for creating visual analytics tools. By connecting machine processing power and capacity, visual analytics is also applicable to huge amounts of heterogeneous and dynamic data, where visualization cannot be used. Several visual analytics tools were developed in the last years. The purpose and effectiveness of those tools varies depending on utilization scenarios, provided visualization techniques and user knowledge. Table 1 presents our comparison of visual analytics tools developed by experts and described in scientific publications. Each row in the table is the main publication corresponding to one visual analytics tool. The references are listed below the table in a chronological order of the publications (top down). Each column of the table corresponds to a subsection from four interrogative questions - why, who, what and how. Each question serves as a category for classifying the tools by:

- Why - Why is it appropriate to use the particular tool, for what reason?
- Who - Who should use or is able to use the tools?
- What - What are the main features that can be analysed and visualized?
- How - How the visualization can be done, what techniques can be used? For this category we distinguish only between three general techniques.

4 Conclusion

This survey is dedicated towards reviewing the state-of-the-art in visualization and visual analytics in data science. We presented both topics separately and compared well-known visual analytics tools based on four categories presented in Sect. 3. Several times we underlined the importance of visualization and visual analytics in data science. Data are more than just numbers and words. Analysing data is similar to storytelling. The stories in this process are dealing with the real world and can be simple and straightforward as well as complex and uncontrollable. We intend to focus our further research on development of a platform for visual data science. Our vision is to allow a flexible and dynamic interconnection of analytical methods and visual techniques with automatic adaptation to data and possible problem statements.

References

1. Thomas, J.J., Cook, K.A.: Illuminating the path. National Visualization and Analytics (2005)
2. Keim, D., Andrienko, G., Fekete, J.-D., Gorg, C., Kohlhammer, J., Melançon, G.: Visual analytics: definition, process, and challenges. In: LNCS, pp. 154–176 (2008)
3. Mansmann, F., Ellis, G.: Mastering the Information Age Solving Problems with Visual Analytics. Taylor & Francis Group, Abingdon (2010)
4. Pohl, M., Smuc, M., Mayr, E.: The user puzzle - explaining the interaction with visual analytics systems. IEEE Trans. Vis. Comput. Graph. **18**(12), 154–176 (2012)

5. Endert, A., Hossain, M.S., Ramakrishnan, N., North, C., Fiaux, P., Andrews, C.: The human is the loop: new directions for visual analytics. J. Intell. Inf. Syst. **43**(3), 411–435 (2014)
6. Sacha, D., Stoffel, A., Stoffel, F., Kwon, B.C., Ellis, G., Keim, D.: Knowledge generation model for visual analytics. IEEE Trans. Vis. Comput. Graph. **20**(12), 1604–1613 (2014)
7. Kwon, B.C., Eysenbach, B., Verma, J., Ng, K., De Filippi, C., Stewart, W.F., Perer, A.: Clustervision: visual supervision of unsupervised clustering. IEEE Trans. Vis. Comput. Graph. **24**(1), 1604–1613 (2018)
8. Stasko, J., Gorg, C., Liu, Z., Singhal, K.: Jigsaw: supporting investigative analysis through interactive visualization. In: IEEE Symposium on Visual Analytics Science and Technology, pp. 131–138 (2007)
9. Cavallo, M., Demiralp, Ç.: Clustrophile 2: guided visual clustering analysis. IEEE Trans. Vis. Comput. Graph. **25**(1), 267–276 (2019)
10. Park, D., Kim, S., Lee, J., Choo, J., Diakopoulos, N., Elmqvist, N.: ConceptVector: text visual analytics via interactive lexicon building using word embedding. IEEE Trans. Vis. Comput. Graph. **24**(1), 361–370 (2018)
11. Gotz, D., Stavropoulos, H.: DecisionFlow: visual analytics for high-dimensional temporal event sequence data. IEEE Trans. Vis. Comput. Graph. **20**(12), 1783–1792 (2014)
12. Höferlin, B., Netzel, R., Höferlin, M., Weiskopf, D., Heidemann, G.: Inter-active learning of ad-hoc classifiers for video visual analytics. In: IEEE Conference on VAST, pp. 23–32 (2012)
13. Jeong, D. H., Ziemkiewicz, C., Fisher, B., Ribarsky, W., Chang, R.: iPCA: an interactive system for PCA-based visual analytics. In: Computer Graphics Forum. (28)3, pp. 767–774. (2009)
14. Brown, E.T., Liu, J., Brodley, C.E., Chang, R.: Dis-function: learning distance functions interactively. In: IEEE Conference on VAST, pp. 83–92 (2012)
15. Alsakran, J., Chen, Y., Zhao, Y., Yang, J., Luo, D.: STREAMIT: dynamic visualization and interactive exploration of text streams. In: IEEE Pacific Visualization Symposium, pp. 131–138 (2011)
16. Yi, J.S., Melton, R., Stasko, J., Jacko, J.A.: Dust & magnet: multivariate information visualization using a magnet metaphor. Inform. Vis. **4**(4), 239–256 (2005)
17. Krause, J., Perer, A., Bertini, E.: INFUSE: interactive feature selection for predictive modeling of high dimensional data. IEEE Trans. Vis. Comput. Graph. **20**(12), 1614–1623 (2014)
18. Strobelt, H., Gehrmann, S., Pfister, H., Rush, A.M.: LSTMVis: a tool for visual analysis of hidden state dynamics in recurrent neural networks. IEEE Trans. Vis. Comput. Graph. **24**(1), 667–676 (2018)
19. Bögl, M., Aigner, W., Filzmoser, P., Lammarsch, T., Miksch, S., Rind, A.: Visual analytics for model selection in time series analysis. IEEE Trans. Vis. Comput. Graph. **19**(12), 2237–2246 (2013)
20. Krause, J., Perer, A., Stavropoulos, H.: Supporting iterative cohort construction with visual temporal queries. IEEE Trans. Vis. Comput. Graph. **22**(1), 91–100 (2016)
21. Chandrasegaran, S., Badam, S.K., Kisselburgh, L., Peppler, K., Elmqvist, N., Ramani, K.: VizScribe: a visual analytics approach to understand designer behavior. Int. J. Hum. Comput. Stud. **100**, 66–80 (2017)

22. Du, F., Plaisant, C., Spring, N., Shneiderman, B.: EventAction: visual analytics for temporal event sequence recommendation. In: IEEE Conference on VAST, pp. 61–70 (2016)
23. Kim, M., Kang, K., Park, D., Choo, J., Elmqvist, N.: TopicLens: efficient multi-level visual topic exploration of large-scale document collections. IEEE Trans. Vis. Comput. Graph. **23** (1), 151–160 (2017)
24. Kwon, B.C., Kim, H., Wall, E., Choo, J., Park, H., Endert, A.: AxiSketcher: interactive nonlinear axis mapping of visualizations through user drawings. IEEE Trans. Vis. Comput. Graph. **23**(1), 221–230 (2017)
25. Pezzotti, N., Höllt, T., Gemert, J.V., Lelieveldt, B.P.F., Eisemann, E., Vilanova, A.: DeepEyes: progressive visual analytics for designing deep neural networks. IEEE Trans. Vis. Comput. Graph. **24**(1), 98–108 (2018)
26. Bailey, S.M., Wei, J.A., Wang, C., Parra, D., Brusilovsky, P.: CNVis: a web-based visual analytics tool for exploring conference navigator data. Electron. Imaging **1**, 1–11 (2018)