On the Challenges and Opportunities in Visualization for Machine Learning & Knowledge Extraction: A Research Agenda

Cagatay Turkay[1], Robert Laramee[2], Andreas Holzinger[3]

[1] giCentre, Department of Computer Science
City, University of London, UK
Cagatay.Turkay.1@city.ac.uk
[2] Department of Computer Science
University of Swansea
r.s.laramee@swansea.ac.uk
[3] Holzinger Group, HCI-KDD, Institute for Medical Informatics/Statistics,
Medical University Graz, Austria
andreas.holzinger@hci-kdd.org

Abstract. We describe a selection of challenges at the intersection of machine learning and data visualization and outline a subjective research agenda based on professional and personal experience. The unprecedented increase in the amount, variety and the value of data has been significantly transforming the way that scientific research is carried out and businesses operate. Within data science, which has emerged as a practice to enable this data-intensive innovation by gathering together and advancing the knowledge from fields such as statistics, machine learning, knowledge extraction, data management, and visualization, visualization plays a unique and maybe *the* ultimate role as an approach to facilitate the human and computer cooperation, and to particularly enable the analysis of diverse and heterogeneous data using complex computational methods where algorithmic results are challenging to interpret and operationalize. Whilst algorithm development is surely at the center of the whole pipeline in disciplines such as Machine Learning and Knowledge Discovery, it is visualization which ultimately makes the results accessible to the end user. Visualization thus can be seen as a mapping from arbitrarily high-dimensional abstract spaces to the lower dimensions and plays a central and critical role in interacting with machine learning algorithms, and particularly in interactive machine learning (iML) with including the human-in-the-loop. The central goal of the CD-MAKE VIS workshop is to spark discussions at this intersection of visualization, machine learning and knowledge discovery and bring together experts from these disciplines. This paper discusses a perspective on the challenges and opportunities in this integration of these discipline and presents a number of directions and strategies for further research.

Keywords: Visualization, Machine Learning, Knowledge Extraction

1 Introduction

The unprecedented increase in the amount, variety and the value of data has been significantly transforming the way that scientific research is carried out and businesses operate. Knowledge generated from data drives innovation in almost all application domains, including health, transport, cyber security, manufacturing, digital services, and also scientific domains such as biology, medicine, environmental and physical sciences, the humanities and social sciences to name a few [1]. The archived data, however, is becoming increasingly complex and heterogeneous, and making sense of such data collections is becoming increasingly challenging.

Algorithmic approaches are increasingly providing effective solutions to problems that are related to well-defined tasks such as classification or predictive modelling based on trend analysis to name a few [2]. However, there are several other problems where the objectives are much less well-defined – often leading to partial or uncertain computational results that require manual interventions from analysts to be useful, or to results that are so complex that interpretation is a barrier against their effective use. The field of visualization, and in particular visual analytics, is a discipline that is motivated by these complex problems that require a concerted effort from computational methods and the human analyst to be addressed [3]. Expert users have the domain knowledge to steer algorithmic power to where it is needed the most [4], and offer the capability and creativity to fine-tune computational results and turn them into *data-informed decisions* [5]. As a growing field, there are already several effective examples where the combination of visualisation and machine learning are being developed to offer novel solutions for data-intensive problems [1].

Visualization of machine learning results will become even more important in the future as with new European regulations, there emerges a need of interpretability of machine learning outcomes, which poses enormous challenges on the visualization, because machine learning techniques for data analysis can be basically seen as a problem of pattern recognition, and there is a (not so little) gap between data modeling and knowledge extraction. Machine learning models may be described in diverse ways, but to consider that some knowledge has been achieved from their description, cognitive factors of the users have to be considered. Even worse, such models can be useless unless they *can be* interpreted, and the process of human interpretation follows rules that go well beyond technical understanding. Consequently, interpretability is a paramount quality that machine learning methods should aim to achieve if they are to be applied for solving practical problems of our daily life [6].

In this position statement, we stress the importance of this merger between these two fields and present a mini-research agenda to spark and inform the discussions at the 2017 CD-MAKE VIS workshop. We present here a non-comprehensive but a representative sample of recent work, discuss opportunities in further research, and challenges in bringing these domains together.

2 A few examples of Visualization and Machine Learning Integration

Visualization is an important method of transforming the symbolic into the geometric, offers opportunities for discovering knowledge in data and fosters insight into data [7]. There are several examples for the importance of visualization in health, e.g. Otasek et al. [8] present work on Visual Data Mining (VDM), which is supported by interactive and scalable network visualization and analysis. Otasek et al. emphasize that knowledge discovery within complex data sets involves many workflows, including accurately representing many formats of source data, merging heterogeneous and distributed data sources, complex database searching, integrating results from multiple computational and mathematical analyses, and effectively visualizing properties and results. Mueller et al. [9] demonstrate the successful application of data Glyphs in a disease analyser for the analysis of big medical data sets with automatic validation of the data mapping, selection of subgroups within histograms and a visual comparison of the value distributions.

A good example for the catenation of visualization with ML is clustering: Clustering is a descriptive task to identify homogeneous groups of data objects based on the dimensions (i.e. values of the attributes). Clustering methods are often subject to other systems, for example to reduce the possibility of recommender systems (e.g. Tag-recommender on Youtube videos [10]); for example clustering of large high-dimensional gene expression data sets has widespread application in -omics [11]. Unfortunately, the underlying structure of these natural data sets is often fuzzy, and the computational identification of data clusters generally requires (human) expert knowledge about cluster number and geometry.

The high-dimensionality of data is a huge problem in health informatics - but in many other domains - and the curse of dimensionality is a critical factor for clustering: With increasing dimensionality the volume of the space increases so fast that the available data becomes sparse, hence it becomes impossible to find reliable clusters; also the concept of distance becomes less precise as the number of dimensions grows, since the distance between any two points in a given data set converges; moreover, different clusters might be found in different sub spaces, so a global filtering of attributes is also not sufficient. Given that large number of attributes, it is likely that some attributes are correlated, therefore clusters might exist in arbitrarily oriented affinity sub spaces. Moreover, high-dimensional data likely includes *irrelevant* features, which may obscure to find the relevant ones, thus increases the danger of modeling artifacts - issues which are simply not true. The problem is that we are confronted with subjective similarity functions; there are a lot of examples of subjective grouping in our daily life (e.g. cars are perceived differently).

Subspace clustering problems are hard, because for the grouping very different characteristics can be used: highly subjective and context specific. What is recognized as comfort for end-users of individual systems, can be applied in scientific research for the interactive exploration of high-dimensional data sets [12].

Consequently, iML-approaches can be beneficial to support finding solutions in hard biomedical problems [13].

Humans are good in comparison for the determination of similarities and dissimilarities - described by nonlinear multidimensional scaling (MDS) models [14]. MDS models represent similarity relations between entities as a geometric model that consists of a set of points within a metric space. The output of an MDS routine is a geometric model of the data, with each object of the data set represented as a point in n-dimensional space. In such operations, the human intervention can help in generating semantically relevant projections or in curating new "bespoke" projection axes [15].

A relatively new technique in that respect is t-SNE [16] that visualizes high-dimensional data by giving each data point a location in a two or three-dimensional map. t-SNE was tested with many different data sets and showed much better results than e.g. Isomap [17] or Locally Linear Embedding [18].

Visualizing properties of t-SNE is discussed by Martin Wattenberg and Fernanda Viegas in their recent keynote talk at the EuroVis 2017 Conference in Baracelone, Spain, "Visualization: The Secret Weapon of Machine Learning.". See Figure 1.

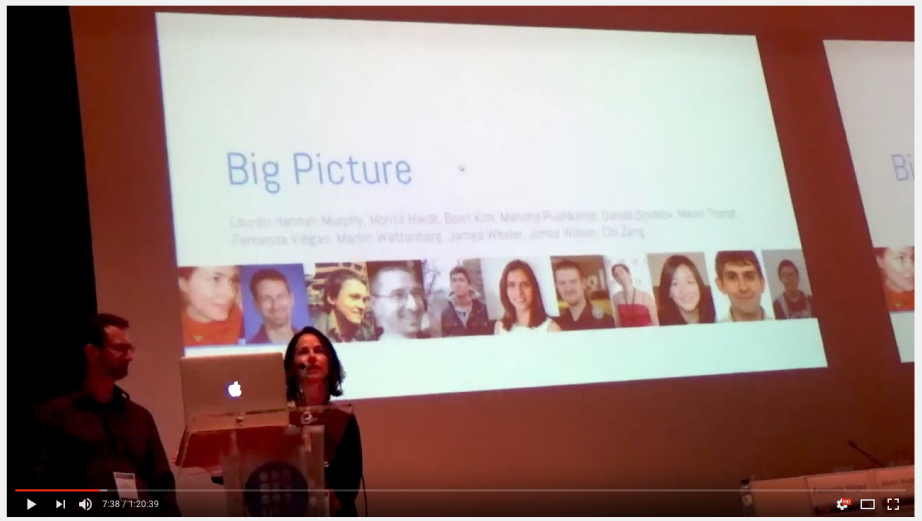

Fig. 1. The Keynote Talk delivered by Martin Wattenberg and Fernanda Viegas at the EuroVis Conference in 2017, Barcelona, Spain. This excellent and informative talk is archived and can be viewed at the following url: `https://youtu.be/E70lG9-HGEM`

The above discussion only scratches the surface in this multidisciplinary research area, for more comprehensive selection of work, good starting points are review papers by Endert et al. [1] and Sacha et al. [19].

3 Challenges and Opportunities for Research

In this section we identify a short list of potential future research directions that can build on a synergy between these disciplines.

Comprehensive analysis of several disjoint data sources: With the advent of new data generation and collection mechanisms, analysts have now the chance to work with not only a single data set but with several data sets coming from diverse channels with different characteristics [20]. In addition to data that is available within their own organisations, they also have access to extremely rich, open data repositories that can add significant value to their analyses. Analysts are often advised to be to be "magnetic" towards new data sources in their everyday practices and integrate a large variety of information sources to generate valuable insight [21]. Analysts will benefit significantly from models that can learn from multiple data sources, however, existing tools are often developed to work with a single, well-defined data source. Making sense of such diverse information sources that are not even physically linked requires methods where "semantic" links between the data sources are built exploratively and visualization can play an important role to facilitate such cross-dataset analysis.

Developing balanced interaction models for human-machine collaboration: Interactive data analysis solutions rely on the effective elicitation of expert knowledge in steering the computational methods whilst dealing with ill-defined problems. However, interaction is a costly operation and experts are often in need of approaches that provide them reliable solutions quickly and accurately [22]. In order to develop effective human-in-the-loop systems [23], there is a need to strike the right balance between computation and human initiative. Certain, often well-defined, tasks are better suited for computation and tasks can be broken down into smaller sub-tasks where little user intervention is needed and user's role is then to harmonize the various observations made through these sub-components. Such an approach can make these systems more effective and utilizes expert's knowledge for tasks where it matters the most. Designing such systems, however, requires an in-depth understanding of tasks, and rigorous user testing.

"Learning" the user: An area where machine learning models can enhance interactive data analysis is through methods that are trained on user interaction data. Users interact with interactive systems in distinctive ways and often the successful execution of an analysis session depends on the accurate identification of *user intent* [1]. Algorithms can be trained on user activity logs to understand the user better and can offer "personalized" analytical recommendations to improve the data analysis process.

Visual storytelling for enhanced interpretation and algorithmic transparency: An emerging trend in visualization is the use of storytelling techniques to communicate concepts effectively [24]. As algorithms get more and more complex

and tend to carry black-box characteristics, interpreting an algorithmic outcome is increasingly gaining importance. Recent examples by the Google Big Picture team[4] or promising innovative publications such as Distill.pub[5] are demonstrating how an effective use of visualization can help unravel algorithms and make them more accessible for wider audiences and also help in educational purposes. However, examples so far are often designed case-by-case basis and further research is needed to develop guidelines, best-practices and a systematic characterization of the role and scope of visualization and interaction.

The list above highlights some of the emerging and core opportunities for joint research projects, however, the field is open for innovation and it is expected to evidence the emergence of new ideas and topics as the discipline matures.

4 A Potential Road-map for Bridging the Communities

Visualization and machine learning communities are currently disjoint communities with limited overlap between the researchers actively contributing to both domains.

There are several recent initiatives, including the yearly CD-MAKE conference, or the MAKE-Journal [25] that aims to bring the two communities closer. The recently organized Dagstuhl events on "*Bridging Information Visualization with Machine Learning*" [26,27] are solid efforts to bring together researchers work on joint projects. There is now a machine learning tutorial at EuroVis[6] conference, however, more presence from the visualization domain within Machine Learning events, such as NIPS[7], ICML[8], or KDD[9] is needed to transfer state-of-the-art research in visualization over to the Machine Learning domain and vice versa. How important this is can be inferred by a recent discussion during the Google's Vision 2016 talks[10], where Google's director of product, Aparna Chennapragada emphasized that *"the UI must be proportional to AI"* .

5 Conclusion

This position paper discussed some of the emerging trends, opportunities, and challenges in the merger of visualization, machine learning and knowledge discovery domains. As evidenced by the increasing activities and recent publications in the area, there is great potential for impactful future research that is likely to transform the ways that data-intensive solutions and services are designed and

[4] https://research.google.com/bigpicture/
[5] http://distill.pub/
[6] http://mlvis2017.hiit.fi/
[7] https://nips.cc/
[8] https://2017.icml.cc/
[9] http://www.kdd.org/
[10] https://www.youtube.com/watch?v=Rnm83GqgqPE

developed. We identify a number of challenges to spark further discussions and research, however, the presented list is far from being comprehensive and reflects a biased overview of the authors. This multidisciplinary research field is open to several other novel problems and developments that can significantly contribute to the societal and academic impact of the existing research carried out in both domains. As it stands now, the biggest barrier to such multidisciplinary research is the limited number of joint venues to *bridge* the two communities and a limited interest in both communities. There is, however, increasing awareness and appreciation of the ongoing research in both fields and it is highly likely that this merger of visualization and machine learning will be attracting further attention.

References

1. Endert, A., Ribarsky, W., Turkay, C., Wong, B.W., Nabney, I., Blanco, I.D., Rossi, F.: The state of the art in integrating machine learning into visual analytics. Computer Graphics Forum (2017)
2. Marsland, S.: Machine learning: an algorithmic perspective. CRC press (2015)
3. Keim, D.A., Mansmann, F., Schneidewind, J., Thomas, J., Ziegler, H.: Visual analytics: Scope and challenges. In: Visual data mining. Springer (2008) 76–90
4. Williams, M., Munzner, T.: Steerable, progressive multidimensional scaling. In: Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on, (IEEE) 57–64
5. Turkay, C., Slingsby, A., Lahtinen, K., Butt, S., Dykes, J.: Supporting theoretically-grounded model building in the social sciences through interactive visualisation. Neurocomputing (2017) –
6. Vellido, A., Martn-Guerrero, J.D., Lisboa, P.J.: Making machine learning models interpretable. In: ESANN - European Symposium on Artificial Neural Network. Volume 12. (2012) 163–172
7. Ward, M., Grinstein, G., Keim, D.: Interactive data visualization: foundations, techniques, and applications. AK Peters, Ltd. (2010)
8. Otasek, D., Pastrello, C., Holzinger, A., Jurisica, I.: Visual data mining: Effective exploration ofthe biological universe. In Holzinger, A., Jurisica, I., eds.: Interactive Knowledge Discovery and Data Mining in Biomedical Informatics: State-of-the-Art and Future Challenges. Lecture Notes in Computer Science LNCS 8401. Springer, Heidelberg, Berlin (2014) in print
9. Mueller, H., Reihs, R., Zatloukal, K., Holzinger, A.: Analysis of biomedical data with multilevel glyphs. BMC Bioinformatics **15** (2014) S5
10. Toderici, G., Aradhye, H., Paca, M., Sbaiz, L., Yagnik, J.: Finding meaning on youtube: Tag recommendation and category discovery. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010), IEEE (2010) 3447–3454
11. Sturm, W., Schreck, T., Holzinger, A., Ullrich, T.: Discovering medical knowledge using visual analytics a survey on methods for systems biology and omics data. In Bühler, K., Linsen, L., John, N.W., eds.: Eurographics Workshop on Visual Computing for Biology and Medicine (2015), Eurographics EG (2015) 71–81
12. Müller, E., Assent, I., Krieger, R., Jansen, T., Seidl, T.: Morpheus: interactive exploration of subspace clustering. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining KDD 08, ACM (2008) 1089–1092

13. Hund, M., Sturm, W., Schreck, T., Ullrich, T., Keim, D., Majnaric, L., Holzinger, A.: Analysis of patient groups and immunization results based on subspace clustering. In Guo, Y., Friston, K., Aldo, F., Hill, S., Peng, H., eds.: Brain Informatics and Health, Lecture Notes in Artificial Intelligence LNAI 9250. Volume 9250. Springer International Publishing, Cham (2015) 358–368
14. Shepard, R.N.: The analysis of proximities: Multidimensional scaling with an unknown distance function. Psychometrika **27** (1962) 125–140
15. Kim, H., Choo, J., Park, H., Endert, A.: Interaxis: Steering scatterplot axes via observation-level interaction. IEEE transactions on visualization and computer graphics **22** (2016) 131–140
16. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9** (2008) 2579–2605
17. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290** (2000) 2319–2323
18. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290** (2000) 2323–2326
19. Sacha, D., Zhang, L., Sedlmair, M., Lee, J.A., Peltonen, J., Weiskopf, D., North, S.C., Keim, D.A.: Visual interaction with dimensionality reduction: A structured literature analysis. IEEE transactions on visualization and computer graphics **23** (2017) 241–250
20. Kehrer, J., Hauser, H.: Visualization and visual analysis of multifaceted scientific data: A survey. IEEE transactions on visualization and computer graphics **19** (2013) 495–513
21. Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J.M., Welton, C.: Mad skills: new analysis practices for big data. Proceedings of the VLDB Endowment **2** (2009) 1481–1492
22. Yi, J.S., ah Kang, Y., Stasko, J.: Toward a deeper understanding of the role of interaction in information visualization. IEEE transactions on visualization and computer graphics **13** (2007) 1224–1231
23. Holzinger, A.: Interactive machine learning for health informatics: When do we need the human-in-the-loop? Springer Brain Informatics (BRIN) **3** (2016) 119–131
24. Kosara, R., Mackinlay, J.: Storytelling: The next step for visualization. Computer **46** (2013) 44–50
25. Holzinger, A.: Introduction to machine learning & knowledge extraction (make). Machine Learning and Knowledge Extraction **1** (2017) 1–20
26. Keim, D.A., Rossi, F., Seidl, T., Verleysen, M., Wrobel, S.: Information Visualization, Visual Data Mining and Machine Learning (Dagstuhl Seminar 12081). Dagstuhl Reports **2** (2012) 58–83
27. Keim, D.A., Munzner, T., Rossi, F., Verleysen, M.: Bridging Information Visualization with Machine Learning (Dagstuhl Seminar 15101). Dagstuhl Reports **5** (2015) 1–27