Data Exploration and Visual Analytics Challenges in AI Era

Appears in ACM SIGMOD Blog
January 2025

The *International Workshop on Big Data Visual Exploration and Analytics* (BigVis) is an annual event, which brings together scholars from the communities of Data Management & Mining, Information Visualization, Machine Learning and Human-Computer Interaction. The 7th BigVis event (BigVis 2024)[1] was organized in conjunction with the *50th International Conference on Very Large Databases* (VLDB 2024) in Guangzhou, China. The organizing committee invited 5 research experts to provide their insights and discuss future research challenges related to Data Exploration and Visual Analytics.

The invited experts are Sihem Amer-Yahia (CNRS, Université Grenoble Alpes), Leilani Battle (University of Washington), Yifan Hu (Northeastern University), Dominik Moritz (Carnegie Mellon University & Apple) and Aditya Parameswaran (Berkeley University). This post presents their responses:

1. Truly Personalized Visual Data Exploration

Sihem Amer-Yahia, CNRS, Université Grenoble Alpes, France

Future Challenges

A visualization is a pair of the form (data, visual element). For instance, data could be a distribution of gender values and the visual element could be a pie chart. A distribution could be mapped to various elements using algebras such as Vega-Lite [1]. Different users appreciate different such mappings. It is hence only natural to ask the question: ***What is the meaning of personalized data visualization when both data and mapping of data to visual elements can be personalized*?** Should they both be personalized? Usually, only data is personalized. A follow up question is: ***Can we learn those***

[1]https://bigvis.imsi.athenarc.gr/bigvis2024

visualizations from observing people interacting with data and visual elements? How expressive should this interaction be? How granular should it be? Should we decouple data recommendations from visual recommendations and allow people to provide feedback on them separately and together? Can we do that in a privacy-preserving manner? Finally, since we are talking about users interacting with data and visualizations, *how to ensure sub-second interaction time in personalized visual exploration*?

These challenges lay the ground for new research that is truly at the crossroad of databases for privacy-preserving machine learning and declarative visualization. A promising avenue to address all three challenges is to combine the relational algebra with Vega-Lite to extend it to handle in-database Multi-Armed Bandits for online learning [2] and use in-database federated learning to ensure privacy [3]. The augmented algebra will provide the ability to blend data and visualization recommendations in the same framework. It will open new opportunities to co-optimize the training and inference of ML models while taking into consideration user feedback. This is possible by revisiting the reward according to the data and visual elements that are best preferred by the user as they explore (data, visual element) pairs in sequence. Speeding up ML inference can be done by materializing views representing the most common exploration pathways and personalizing them on-the-fly. Finally, federated learning can leverage cryptographic schemes such as homomorphic encryption to compute rewards without revealing data in clear. That is particularly important when multiple users explore data.

Emerging Applications

The framework outlined above is applicable in a key emerging application, enabling visual exploratory data analysis for users with different data science and domain expertise. We developed *Dora The Explorer*, an RL-powered data exploration system for astronomers [4], and DashBot [5], an adaptive MAB-powered system for medical professionals. The astronomers we worked with were well-versed in data science and preferred partially guided approaches where they could intervene and change the automatic decisions made by the machine learned models. The doctors we worked demanded to validate the patient analytics that were recommended to them. This was fed back into the system to best learn which next analytics to provide. *Enabling this conversation between automated data exploration and people with different expertise levels is the next frontier* in personalized visual data exploration.

2. Effective Guidance and Interoperability

Leilani Battle, University of Washington, USA

How do we provide the "right" guidance?

With the rise of AI assistants in data science [9], we are seeing more and more auto-generated recommendations for what analysts should explore next in their data [13]. However, how do we know that these assistants are providing the right guidance [10]? For example, current techniques often

produce generic recommendations with limited benefits over standard baselines [13][6], which may be due in part to the underlying algorithms and models lacking *domain context* when generating recommendations [6]. Further, certain analysts may be too trusting of auto-generated recommendations regardless of rigor/quality [12]. Given their potential influence over human analysts, how can we ensure that our recommendations avoid introducing bias or causing harm? We need new methods for incorporating domain context into the underlying algorithms and AI models as well as benchmarks for validating their outputs. Furthermore, we need a deeper understanding of how auto-generated recommendations influence the reasoning of human analysts [9]. As part of this effort, we encourage the community to expand and strengthen collaborations across data science (e.g., among HCI, visualization, and data management researchers) as well as with other fields such as psychology and cognitive science.

How do we optimize interoperable components within larger data science ecosystems?

Computational notebooks are integral to collaborative data science work given their flexibility and ease of sharing with others [11][14]. However, notebook environments like JupyterLab can be challenging to design for [8]. Further, the data management and data visualization communities are used to building systems they control [7]. Notebooks contradict that assumption. Beyond the known challenges of working within an existing development environment, we observe few projects that profile or optimize the performance of data management/visualization tools that run alongside them. For example, the user drives the workload that big data researchers intend to optimize, but the user's interactions with a DBMS may be influenced by other tools and packages they have also imported into the underlying notebook. Thus, we must investigate how this blending of tools may alter the typical data processing workloads we see in other big data contexts. Otherwise, our optimizations may miss the full context of how a tool is being used and thus fail to address the user's performance concerns.

3. Auto-generated Big Network Data Visualization and Story Telling by Learning Human Sense of Aesthetics

Yifan Hu Northeastern University, USA

Automatically learning to generate aesthetically pleasing networking visualization through human examples

Information visualization enables the graphical representation of data and information that helps the users to find patterns, trends, and relationships within the data, and to communicate complex data sets clearly and effectively. In the field of network visualization, traditionally, visualization of undirected graphs relies on heuristics, especially those based on modeling graphs as physical systems, with the belief that minimizing the energy of such systems lead to aesthetic looking layout of the graph that

help to illustrate the underlining unstructured data. There are two problems with this approach. First, traditional force directed algorithms [15][16][17] can have high complexity and may not find the optimal solution. Second, and more importantly, it is not proven that human aesthetic preference can be modeled well by physical systems in all cases. With the advent of deep learning, there is a growing interest in leveraging the power of neural networks to help network visualization (e.g., [18][19]). In particular, it is now possible to use deep learning to expand the horizon of traditional graph visualization in a number of directions. Firstly, it was demonstrated that a graph neural network (GNN) model is able to be employed to optimize arbitrary differentiable objective functions. Once trained, such a model can be applied to arbitrary graphs never seen in the training data and create visualizations which optimize the objective function even better than traditional force directed algorithms [20][21][23][23]. Secondly, it was shown that by using Generative Adversarial Network (GAN), we can even train neural networks to optimize non-smooth objective functions [24]. In fact, this approach does not require the access of the objective function at all, and only requires a comparative function that can choose between two visualizations, the "better" one (based on a hidden objective function unknown to GAN). This opens the avenues for future investigation: instead of optimizing the energy of the physical system, or other arbitrary objectives such as edge crossing, we should be able to model the human sense of aesthetics directly. For example, if reducing edge cross is what's most important for human, Tiezzi et al. [21] demonstrated that a fully connected network can be used to classify whether two edges cross; on the other hand, Wang et al [24] has shown that the discriminator in a GAN setup can be trained to choose drawings with lower number of edge crossing, by "teaching" it only with pairs of bad and good drawings examples. The holy grail is to train a model on a collection of example visualizations, to learn human visual preference. We believe such a system is possible, but the challenges are to be able to scale GNN, GAN or other deep learning-based models to very large unstructured data, as well as to collect large amounts of human preference examples. We have witnessed some progress in the scalability area already, e.g., by the use of multilevel paradigm [25][26][27]. Further work is needed to curate large amounts of training data, and to make deep learning-based network visualization adapt to human preferences, incorporating human specified constraints, and to run even faster than traditional network layout algorithms.

Generate not just the visualization but also the story behind it

Another future direction for research involves not just producing the visualization itself but also crafting a narrative, or even combining it with animations/videos that elucidate the patterns depicted within the visualization, thus directing the user's attention to the most crucial aspects of generated visualization. Accomplishing this task will necessitate leveraging a multimodal large language model (LLM) trained extensively on a diverse corpus consisting of source data (e.g., CSV files, unstructured data), their visualizations, and their corresponding captions extracted from past visualization and other scientific literature. If we can solve the above challenges, we can then apply such a model to any data to be

analyzed, and automatically generate sample visualizations, animations and narratives that are not only aesthetically pleasing to look at, but also informative, and with a well told story about the visualization.

4. Queries at the Speed of Human Thought

Dominik Moritz, Carnegie Mellon University & Apple, USA

Machine learning development has shifted from being primarily model-centric to primarily data-centric [28]. In the early days of ML, engineers often trained models on hundreds or thousands of data points and meticulously tweaked the modeling function so as not to overfit the data. With the popularization of deep learning, however, data became increasingly important, and many models were derived from pre-trained models. Today, foundation models have similar transformer architectures, and big quality differences stem from the data they are trained or fine-tuned on. The data management community is well-equipped to tackle many of the quality and quantity challenges [29]. Yet, the full AIML lifecycle of requirement elicitation [30], data preparation, monitoring, tuning, and evaluation requires oversight by people who have to slice and dice millions or billions of data records. Since models synthesize patterns, a single record rarely defines the behavior of a model, and therefore ML engineers need to understand patterns and trends in relevant subsets [31]. Relevant subsets are hard to predict and can rarely be precomputed. ML engineers need to grasp often subtle patterns in large datasets full of information. As Herbert A. Simon puts it *"What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently"*. This need to efficiently allocate attention is where data visualization becomes crucial.

Data visualization leverages our powerful perceptual systems to let us see patterns and trends in data. Data visualization in the AIML lifecycle gives us a way to overview and design datasets for training, fine-tuning, and evaluation. It enables serendipitous discovery of data patterns and issues. At the core of the necessary interactive analysis are interfaces that are fast and fast to use. On the one hand, well-designed mixed initiative systems [32] for analysis make people faster [33][34]. We should continue to invest in good tooling that eliminates barriers to effective analysis. On the other hand, delays in interactive systems lead to fewer observations made [35] and could steer analysts towards convenient and fast data along with all the resulting biases. Fast databases, approximation [36][37], prefetching [38], and indexing techniques [39][40] can help developers build fast interactive systems. However, they add complexity for tool builders that may prevent them from quickly adapting interfaces to new needs. Modern data architectures abstract from the low-level optimizations like Mosaic [41] and are being adopted rapidly. Yet, many opportunities remain for deeper integration with databases, which are often optimized for order-of-second responses rather than real-time order-of-millisecond responses or may not support many of the encoding transformations used in effective data visualizations like cartographic

projections. A lot of exciting work remains to explore new ways to effectively present billion-record datasets at the speed of human thought to facilitate effective analysis and communication.

5. Harnessing Generative AI for Data Analytics

Aditya Parameswaran, Berkley University, USA

Recent advances in Generative AI, specifically Large Language Models (LLMs), have the promise to fully democratize data work. We have within our grasp the ability to allow anyone, even without a coding background, to do data work at scale—spanning data exploration, transformation, and insight discovery. LLMs can help bridge the gap between users and their data: they can help communicate with users better in natural language, and also understand the data itself better, spanning both structured and unstructured data formats.

Despite their promise, however, LLMs are unfortunately too brittle for data work. They often make mistakes, disregard instructions, and are prone to "hallucinating" incorrect facts. Users find it difficult to both understand the reasoning process of the LLM and recover from LLM failures. Users struggle to even understand how to ask the LLM to perform a specific task: small changes in wording can lead to drastically different outcomes. So, to democratize data work with LLMs, we will need to leverage ideas from multiple disciplines–including databases and human-computer interaction–by tackling the following research questions:

How can we make LLM-powered data tools more robust? The main difficulty with LLMs is that they make mistakes in an unpredictable fashion. Thus, for any data task at scale, it is virtually guaranteed that there will be LLM mistakes. So, to enhance the robustness of LLMs for data tasks, there are several unanswered questions. How do we best (and automatically) decompose tasks into smaller "tightly scoped" ones that are less difficult for LLMs? How do we ensure that these smaller tasks have a certain degree of reliability, perhaps by combining the results of one LLM with those of others with different capabilities? How do we deal with LLM inconsistencies (e.g., LLMs say A>B, B>C and C>A)? Can we catch LLM mistakes before they occur, perhaps by having LLMs themselves synthesize assertions and verify each other's work [44][45]? This work is closely related to the data management community's work on crowdsourcing, which dealt with error-prone humans, much like error-prone LLMs [43].

How can LLM-powered data tools help analyze unstructured data? The vast majority of data exists in unstructured, difficult to analyze, formats including PDFs, word files, images, and videos. Now, with the power of LLMs, we may now be able to empower users to make sense of such datasets via intuitive interfaces [42]. For example, journalists investigating police misconduct may want to extract officer names from internal affairs reports, or search for specific activities in camera feeds, all without code. This leads to several questions, such as: How do we design lightweight interaction modalities that allow

users to "guide" the system by navigating, searching, and highlighting points of interest in these media? How can we best leverage LLMs to automatically craft and suggest transformations based on the content, user guidance, and historical data? How do we design presentation techniques to help users "decide" which among a list of ranked suggestions is the one they would like to pursue?

How should LLM-powered data tools interact with users? Present day LLMs employ a chat-based natural language interface. While this interface is flexible, it is a poor fit for data work as it violates many fundamental human-centered design principles, including giving users a place to start or recover from errors, as well as showcasing the capabilities of the underlying system. Moreover, it sits separate from existing popular data tools such as spreadsheets or computational notebooks, where data exploration and visualization actually happen [47]. So, we ask the question: how should next-generation data work interfaces look like? What are the strengths and weaknesses of various forms of user input, including traditional code, natural language, form-based interfaces, demonstration, examples, and direct manipulation [46], for the purpose of data work? If we do support other forms of input beyond natural language, how do we translate such input to and from natural language so that it can be best interpreted by an LLM? Finally, data work is rarely a solo activity—how do we best support collaboration and handing-off work across individuals?

Overall, I believe this ambitious, but risky vision of democratizing big data work by leveraging the power of LLMs to both understand users and our data better, has the potential to have tremendous impact across a wide variety of domains.

Reference

[1] Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, Jeffrey Heer, Vega-Lite: A Grammar of Interactive Graphics, IEEE TVCG, 2017

[2] Radu Ciucanu, Marta Soare, Sihem Amer-Yahia, Implementing Linear Bandits in Off-the-Shelf SQLite, EDBT 2022

[3] Sotirios Tzamaras, Radu Ciucanu, Marta Soare, Sihem Amer-Yahia, FeReD: Federated Reinforcement Learning in the DBMS, ACM CIKM, 2022

[4] Aurélien Personnaz, Sihem Amer-Yahia, Laure Berti-'Equille, Maximilian Fabricius, Srividya Subramanian. DORA THE EXPLORER: Exploring Very Large Data with Interactive Deep Reinforcement Learning, ACM CIKM, 2021

[5] Sandrine Da Col, , Radu Ciucanu, Marta Soare, Nassim Bouarour, Sihem Amer-Yahia. DashBot: An ML-Guided Dashboard Generation System, ACM CIKM, 2021

[6] Bao, C.S., Li, S., Flores, S.G., Correll, M. and Battle, L., Recommendations for visualization recommendations: Exploring preferences and priorities in public health. ACM *CHI 2022*

[7] Battle, L. and Scheidegger, C. A structured review of data management technology for interactive visualization and analysis. *IEEE TVCG 2020*

[8] Chattopadhyay, S., Prasad, I., Henley, A.Z., Sarma, A. and Barik, T. What's wrong with computational notebooks? Pain points, needs, and design opportunities. ACM CHI, 2020
[9] Gu, K., Grunde-McLaughlin, M., McNutt, A., Heer, J. and Althoff, T. How do data analysts respond to ai assistance? a wizard-of-oz study. ACM CHI, 2024
[10] Heer, J. Agency plus automation: Designing artificial intelligence into interactive systems. Proceedings of the National Academy of Sciences, 2019
[11] Roy, A., Raghunandan, D., Elmqvist, N. and Battle, L. How I Met Your Data Science Team: A Tale of Effective Communication. IEEE VL/HCC, 2023
[12] Zehrung, R., Singhal, A., Correll, M. and Battle, L. Vis ex machina: An analysis of trust in human versus algorithmically generated visualization recommendations. ACM *CHI*, 2021
[13] Zeng, Z., Moh, P., Du, F., Hoffswell, J., Lee, T.Y., Malik, S., Koh, E. and Battle, L. An evaluation-focused framework for visualization recommendation algorithms. IEEE TVCG, 2021
[14] Zhang, A.X., Muller, M. and Wang, D. How do data science workers collaborate? roles, workflows, and tools. ACM *CHI*, 2020
[15] P. Eades. A heuristic for graph drawing. Congressus Numerantium, 1984
[16] Tomihisa Kamada and Satoru Kawai. An algorithm for drawing general undirected graphs. Information Processing Letters, 1989
[17] Emden Gansner, Yehuda Koren and Stephen North, Graph Drawing by Stress Majorization, Graph Drawing, 2004
[18] Y. Wang, Z. Jin, Q. Wang, W. Cui, T. Ma, and H. Qu. Deepdrawing: A deep learning approach to graph drawing. IEEE TVCG, 2019
[19] O.-H. Kwon and K.-L. Ma. A deep generative model for graph layout. IEEE TVCG, 2020
[20] Xiaoqi Wang, Kevin Yen, Yifan Hu, Han-wei Shen, DeepGD: A Deep Learning Framework for Graph Drawing Using GNN, IEEE CG&A, 2021
[21] Matteo Tiezzi, Gabriele Ciravegna, and Marco Gori. Graph neural networks for graph drawing. IEEE Transactions on Neural Networks and Learning Systems, 2024
[22] Yong Wang, Zhihua Jin, Qianwen Wang, Weiwei Cui, Tengfei Ma, and Huamin Qu. Deepdrawing: A deep learning approach to graph drawing. IEEE TVCG, 2020
[23] Loann Giovannangeli, Frederic Lalanne, David Auber, Romain Giot, and Romain Bourqui. Toward efficient deep learning for graph drawing (DL4GD), IEEE TVCG, 2024
[24] Xiaoqi Wang, Kevin Yen, Yifan Hu, and Han-Wei Shen, SmartGD: A GAN-Based Graph Drawing Framework for Diverse Aesthetic Goals, IEEE TVCG, 2023
[25] Kai YAN, Tiejun ZHAO, and Muyun YANG. Graphuly: Graph u-nets-based multi-level graph layout. IEICE Transactions on Information and Systems, 2022
[26] Florian Grötschla and Joël Mathys and Robert Veres and Roger Wattenhofer, CoRe-GD: A Hierarchical Framework for Scalable Graph Visualization with GNNs, International Conference on Learning Representations, 2024
[27] Zengfeng Huang, Shengzhong Zhang, Chong Xi, Tang Liu, and Min Zhou. Scaling up graph neural networks via graph coarsening. ACM SIGKDD, 2021
[28] Zha, Daochen, et al. "Data-centric artificial intelligence: A survey." arXiv preprint, 2023
[29] Whang, Steven Euijong et al. “Data collection and quality challenges in deep learning: a data-centric AI perspective.” VLDBJ, 2021

[30] Robertson, Samantha, et al. "Angler: Helping machine translation practitioners prioritize model improvements." ACM CHI, 2023
[31] Cabrera, Ángel Alexander, et al. "Zeno: An interactive framework for behavioral evaluation of machine learning." ACM CHI, 2023
[32] Allen, James E., Curry I. Guinn, and Eric Horvtz. "Mixed-initiative interaction." IEEE Intelligent Systems and their Applications, 1999
[33] Wongsuphasawat, Kanit, et al. "Voyager: Exploratory analysis via faceted browsing of visualization recommendations." IEEE TVCG, 2015
[34] Epperson, Will, et al. "Dead or alive: Continuous data profiling for interactive data science." IEEE TVCG, 2023
[35] Liu, Zhicheng, and Jeffrey Heer. "The effects of interactive latency on exploratory visual analysis." IEEE TVCG, 2014
[36] Hellerstein, Joseph M., Peter J. Haas, and Helen J. Wang. "Online aggregation". ACM SIGMOD, 1997
[37] Moritz, Dominik, et al. "Trust, but verify: Optimistic visualizations of approximate queries for exploring big data." ACM CHI, 2017
[38] Battle, Leilani, Remco Chang, and Michael Stonebraker. "Dynamic prefetching of data tiles for interactive visualization." ACM SIGMOD, 2016
[39] Liu, Zhicheng, Biye Jiang, and Jeffrey Heer. "imMens: Real-time visual querying of big data." CGF, 2013.
[40] Moritz, Dominik, Bill Howe, and Jeffrey Heer. "Falcon: Balancing interactive latency and resolution sensitivity for scalable linked visualizations." ACM CHI, 2019
[41] Heer Jeffrey, and Dominik Moritz. "Mosaic: An architecture for scalable & interoperable data views." IEEE TVCG, 2023
[42] Lin, Y., Hulsebos, M., Ma, R., Shankar, S., Zeigham, S., Parameswaran, A., & Wu, E. Towards Accurate and Efficient Document Analytics with Large Language Models. arXiv preprint, 2024
[43] Parameswaran, A., Shankar, S., Asawa, P., Jain, N., & Wang, Y. Revisiting prompt engineering via declarative crowdsourcing. arXiv preprint, 2023
[44] Shankar, S., Li, H., Asawa, P., Hulsebos, M., Lin, Y., Zamfirescu-Pereira, J., Chase, H., Fu-Hinthorn, W., Parameswaran, A., & Wu, E. Spade: Synthesizing assertions for large language model pipelines. arXiv preprint, 2024.
[45] Shankar, S., Zamfirescu-Pereira, J., Hartmann, B., Parameswaran, A., & Arawjo, I. Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences. arXiv preprint, 2024
[46] Siddiqui, T., Luh, P., Wang, Z., Karahalios, K., & Parameswaran, A. Shapesearch: A flexible and efficient system for shape-based exploration of trendlines. ACM SIGMOD, 2020
[47] Lee, D.L., Tang, D., Agarwal, K., Boonmark, T., Chen, C., Kang, J., Mukhopadhyay, U., Song, J., Yong, M., Hearst, M., et al. Lux: always-on visualization recommendations for exploratory dataframe workflows. PVLDB, 2021