Visualizing Differences: Interactive Exploration and Comparison of High-Dimensional Data Sets

Klaus Eckelt

PhD Colloquium Computer Science, March 2023

Thesis carried out at the Institute of Computer Graphics,
supervised by Univ.-Prof. Dr. Marc Streit

Thematic Summary

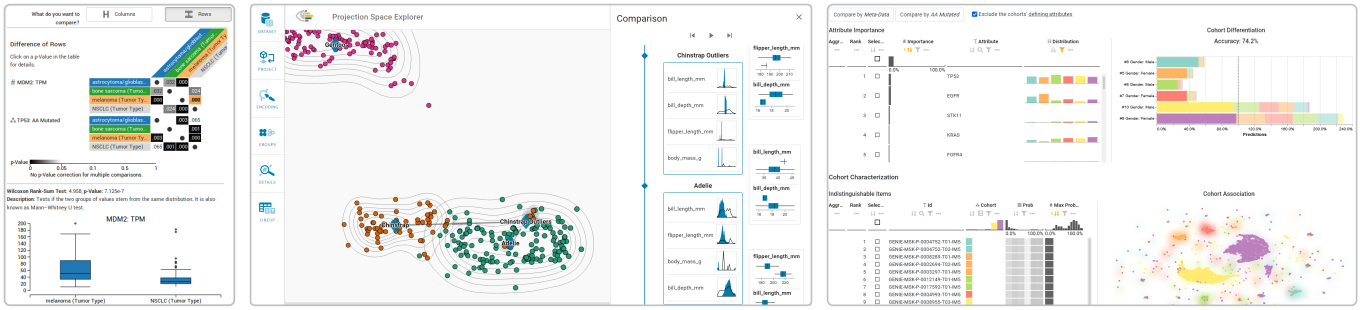
Using data to gain new insights has spread to all areas of business and science. Visualization plays a crucial role in this process, offering a range of techniques for representing data in ways that make it easier to explore and interpret. Interactive visual exploration enables users to generate hypotheses based on visual patterns, compare and identify data sets, and analyze the underlying structure and relationships. Knowing how data differ and what the key differences are is crucial in many domains, especially healthcare, where understanding differences in patient outcomes can lead to better treatments.

My research focuses on techniques for exploring and comparing high-dimensional data sets, particularly cancer patient cohorts. However, these techniques can be applied to tabular data from any domain. Together with collaborators from industry and Johannes Kepler University Linz, I have worked on a series of publications that were accepted and presented at leading journals and conferences in the field of data visualization. The techniques range from comparing groups with statistical tests for individual features (Section A), to visualizing and comparing groups in low-dimensional embeddings of data (Section B), to finding driving differences between groups and characterizing their homogeneity (Section C). These techniques have been incorporated into dedicated visual analytics tools, but could also be used in computational notebooks, which have become increasingly popular [4]. While working in computational notebooks, however, it is often difficult to identify the differences between states. I am therefore currently working on highlighting differences between states of computational notebooks to support iterative data science processes.

A – TourDino: Visualizing Statistical Differences

Seeking relationships and patterns in tabular data is a common data exploration task. To confirm hypotheses that are based on observed visual patterns, users need to switch between exploratory and confirmatory analysis, to compare data sets, and get further information on the significance of the result and the statistical test applied [2].

With TourDino [A], we help users to verify generated hypotheses and confirm insights gained during exploration. TourDino provides several well established methods to compare item groups and features and test the generated hypotheses. On demand, users see details about the applied statistical tests and a visualization to explain the result (see Figure 1a). We have integrated TourDino into two analysis tools for cancer genomics data: Coral [D] and Ordino [7].



(a) TourDino (b) Embedding-Based Group Comparison (c) Kokiri

Figure 1: Techniques to compare high-dimensional data sets discussed in Sections A-C.

B – Embedding-Based Group Visualization and Comparison

While TourDino compares data statistically, differences are often not found within a single feature, but in a larger subset of the high-dimensional data. The challenge of making high-dimensional data accessible for visualizations in a two-dimensional space is typically addressed by dimensionality reduction. To effectively analyze structured data after dimensionality reduction, users need to be able to relate visual patterns to the underlying structure and high-dimensional data.

We have therefore created an exploration approach for scatterplots of low-dimensionally embedded, multivariate data, augmented with structural information about the dataset (see Figure 1b) [B]. Users can define groups in datasets on the fly, compare these groups, and introduce new relationships between them. Summary visualizations show the high-dimensional data for the groups, while difference visualizations show how groups are different from each other. We carefully tailored these summary and difference visualizations to various data types and semantic contexts and integrated the approach in the Projection Space Explorer [1] application.

C – Kokiri: Finding High-Dimensional Differences

In our previous work, we visualized high-dimensional data with structural information in augmented scatterplots by applying dimensionality reduction algorithms. However, we recognized that our difference visualizations were limited in explaining the differences visible in the scatterplots, as they only looked at features in isolation, ignoring combinatorial effects.

Kokiri [C] allows users to compare cohorts by their high dimensional data with the goals of (i) uncovering driving differences between them and (ii) characterizing the homogeneity of individual cohorts. We achieve this by training a random forest model to classify the cohorts based on their high-dimensional data. We report the most important features to differentiate between the cohorts and give an overview of the cohorts' separability . Additionally, the homogeneity of individual cohorts can be assessed based on the classifier's confidence, and hard-to-classify items can be identified. We integrated Kokiri into Coral [D] and demonstrated the utility through a use case comparing lung cancer patient cohorts. The analysis verified previous findings from literature and provided deeper insights into the dataset, uncovering further differences between the patient groups.

Future Work: Visualizing Differences in Iterative Data Science Processes

The previously discussed techniques have been integrated in standalone tools, but dedicated visual analytics tools often lack the flexibility required to handle the complex and diverse data science processes that analysts face [6]. In contrast, computational notebooks, like Jupyter, are highly flexible programming environments that combine code, output, and documentation in a single document, providing a more versatile solution for data science projects [4]. Although computational notebooks should promote reproducibility, prior research has shown that this is not the norm [3]. The iterative process of data science—obtaining, cleaning, profiling, analyzing, and interpreting data—leads to poor coding practices in computational notebooks and non-reproducible results [5].

I am now working on a technique to visually support such iterative and exploratory data science processes. The technique tracks provenance information and visualizes differences between states of the computational notebook, enabling analysts to gain a better understanding of the impact of their changes and the notebook's history. The difference visualization adapts to various types of content in the notebook, such as code, markdown, dataframes, visualizations, or images.

Publications, Talks and Visits

Peer-reviewed Journal and Conference Articles

[A] **K. Eckelt**, P. Adelberger, T. Zichner, A. Wernitznig, and M. Streit. "TourDino: A Support View for Confirming Patterns in Tabular Data". In: *EuroVis Workshop on Visual Analytics (EuroVA '19)* (2019). DOI: 10.2312/eurova.20191117.

[B] **K. Eckelt**, A. Hinterreiter, P. Adelberger, C. Walchshofer, V. Dhanoa, C. Humer, M. Heckmann, C. A. Steinparz, and M. Streit. "Visual Exploration of Relationships and Structure in Low-Dimensional Embeddings". In: *IEEE Transactions on Visualization and Computer Graphics* (2022). DOI: 10.1109/TVCG.2022.3156760.

[C] **K. Eckelt**, P. Adelberger, M. J. Bauer, T. Zichner, and M. Streit. "Kokiri: Random-Forest-Based Comparison and Characterization of Cohorts". In: *IEEE VIS Workshop on Visualization in Biomedical AI* (2022). DOI: 10.1101/2022.08.16.503622.

[D] P. Adelberger, **K. Eckelt**, M. J. Bauer, M. Streit, C. Haslinger, and T. Zichner. "Coral: a web-based visual analysis tool for creating and characterizing cohorts". In: *Bioinformatics* 37.23 (2021), pp. 4559–4561. DOI: 10.1093/bioinformatics/btab695.

Talks and Presentations

2022-10-21 Paper Presentation at the IEEE VIS 2022 Conference
Visual Exploration of Relationships and Structure in Low-Dimensional Embeddings
Oklahoma City, OK, USA

2022-10-17 Paper Presentation at the IEEE VIS 2022 Workshop on Visualization in Biomedical AI
Kokiri: Random-Forest-Based Comparison and Characterization of Cohorts
Oklahoma City, OK, USA

2022-07-13 Invited Highlight Talk at the BioVis@ISMB 2022 Conference
Visual Exploration of Relationships and Structure in Low-Dimensional Embeddings
Madison, WI, USA

2019-06-03 Paper Presentation at the EuroVis Workshop on Visual Analytics 2019
TourDino: A Support View for Confirming Patterns in Tabular Data
Porto, Portugal

Teaching

2020– Explainable AI, Johannes Kepler University Linz, Austria
2018– Visual Analytics, Johannes Kepler University Linz, Austria
2019–2022 Information Visualization, Johannes Kepler University Linz, Austria
2019–2021 Visualisation, imperial College Business School, London, United Kingdom
2018–2022 Computer Graphics, Johannes Kepler University Linz, Austria

I have also supervised several students in their practical work, bachelor's, or master's thesis.

Scientific Community Services

International Program Committee Member

2023 Leipzig Symposium on Visualization In Applications (LEVIA)
2021 - 2022 EuroVis Conference – Poster Track

References

[1] A. Hinterreiter, C. Steinparz, M. Schöfl, H. Stitz, and M. Streit. "Projection Path Explorer: Exploring Visual Patterns in Projected Decision-Making Paths". In: *ACM Transactions on Interactive Intelligent Systems* 11.3–4 (2021), Article 22. DOI: 10.1145/3387165.

[2] D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler. "Challenges in Visual Data Analysis". In: *Proceedings of the Conference on Information Visualisation (InfoVis '06)*. 2006, pp. 9–14. DOI: 10.1109/IV.2006.31.

[3] J. F. Pimentel, L. Murta, V. Braganholo, and J. Freire. "A Large-Scale Study About Quality and Reproducibility of Jupyter Notebooks". In: *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*. 2019, pp. 507–517. DOI: 10.1109/MSR.2019.00077.

[4] F. Psallidas, Y. Zhu, B. Karlas, J. Henkel, M. Interlandi, S. Krishnan, B. Kroth, V. Emani, W. Wu, C. Zhang, M. Weimer, A. Floratou, C. Curino, and K. Karanasos. "Data Science Through the Looking Glass: Analysis of Millions of GitHub Notebooks and ML.NET Pipelines". In: *ACM SIGMOD Record* 51.2 (2022), pp. 30–37. DOI: 10.1145/3552490.3552496.

[5] A. Rule, A. Tabard, and J. D. Hollan. "Exploration and Explanation in Computational Notebooks". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Association for Computing Machinery, 2018, pp. 1–12. DOI: 10.1145/3173574.3173606.

[6] J. Schmidt and T. Ortner. "Visualization in Notebook-Style Interfaces". In: *Proceedings of the Workshop on the Gap between Visualization Research and Visualization Software (VisGap)*. 2020. DOI: 10.2312/visgap.20201104.

[7] M. Streit, S. Gratzl, H. Stitz, A. Wernitznig, T. Zichner, and C. Haslinger. "Ordino: a visual cancer analysis tool for ranking and exploring genes, cell lines and tissue samples". In: *Bioinformatics* 35.17 (2019), pp. 3140–3142. DOI: 10.1093/bioinformatics/btz009.