EuroVis Workshop on Visual Analytics (2019)
C. Turkay and T. von Landesberger (Editors)

TourDino: A Support View for Confirming Patterns in Tabular Data

K. Eckelt[1], P. Adelberger[1], T. Zichner[2], A. Wernitznig[2], and M. Streit[1]

[1]Johannes Kepler University Linz, Institute of Computer Graphics, Linz, Austria
[2] Boehringer Ingelheim RCV GmbH & Co KG, Department of Pharmacology and Translational Research, Vienna, Austria

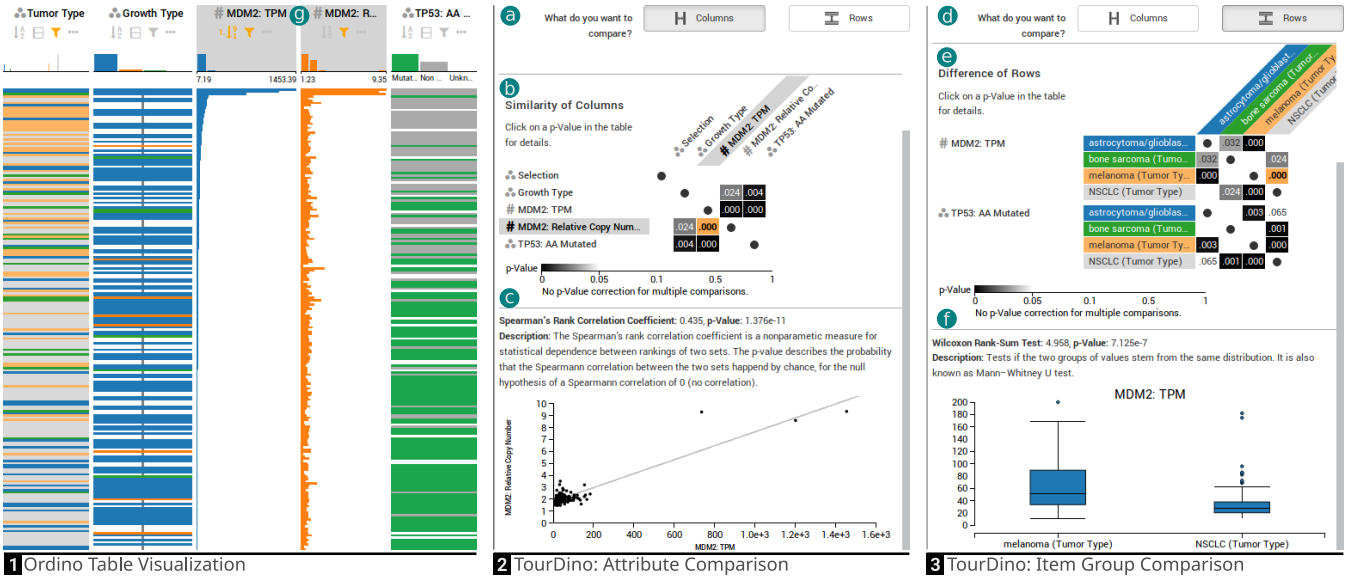

Figure 1: *Ordino [SGS*19], on the left, showing the tabular data in overview mode (1) with two attributes highlighted (g), and two TourDino support views (2,3) on the right: for attribute and item comparison. The support views show the task chooser (a,d), the significance matrix (b,e), and the detail visualization including a brief description of the statistical test applied (c,f).*

Abstract

Seeking relationships and patterns in tabular data is a common data exploration task. To confirm hypotheses that are based on visual patterns observed during exploratory data analysis, users need to be able to quickly compare data subsets, and get further information on the significance of the result and the statistical test applied. Existing tools, however, either focus on the comparison of a single data type, such as comparing numerical attributes only, or provide little or no statistical evaluation to assess a hypothesis. To fill this gap, we present TourDino, a support view that helps users who are not experts in statistics to verify generated hypotheses and confirm insights gained during the exploration of tabular data. In TourDino we present an overview of the statistical significance of various row or column comparisons. On demand, we show further details, including the test score, a textual description, and a detail visualization explaining the results. To demonstrate the efficacy of our approach, we have integrated TourDino in the Ordino drug discovery platform for the purpose of identifying new drug targets.

1. Introduction

Visual exploration is a common way of gaining new insights from tabular data. As we know from well-known examples such as Anscombe's quartet, relying on descriptive statistics is often insufficient to capture the characteristics of multi-dimensional data [MF17]. In order to be able to trust the patterns users observe in a visualization, they need to be confirmed using statistical tests. After all, similar visual patterns can lead to different statistical results. Although users may be domain experts who know the data very well, confirming visual findings is challenging. Which statistical test is appropriate depends on the data type, the tests' assumptions (e.g., a normal distribution), and the hypothesis. Additionally, users may lack the statistical knowledge to understand and trust the results presented.

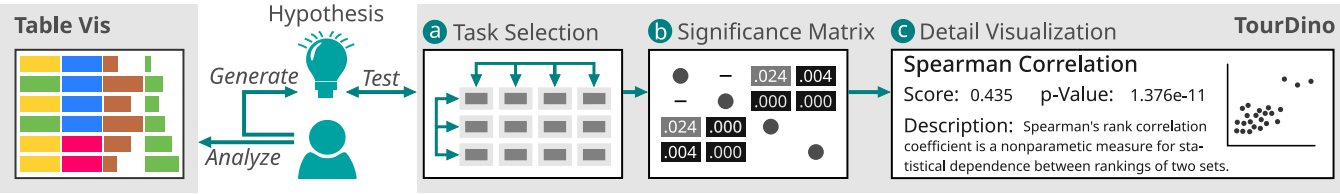

Figure 2: *TourDino tests hypotheses generated in the exploratory analysis. After selecting the comparison task and the data to compare (a), the significance matrix (b) shows all findings. Selecting a cell opens a detail visualization with further details about the statistical test (c).*

The goal of TourDino is to support users in the process of switching between exploratory and confirmatory analysis [KMSZ06]. As an addition to existing table visualizations, TourDino provides several well established methods to compare item groups and attributes and test the generated hypotheses. We show the similarities and dissimilarities found within the data by their significance (see Figure 2b), a value comparable across all methods. On demand, users see details about the applied statistical tests, with a small visualization to explain the result (see Figure 2c).

Throughout this paper, we refer to the result of any statistical test as a *score*. We call the columns of a table *attributes*, and the rows *items* [Mun14, p. 25]. All values of an attribute share the same type and can be either numerical or categorical. Following the definition by Munzner [Mun14, p. 56-57], we aim to find correlations between numerical attributes, and dependencies of one attribute on another categorical attribute. Between groups of items, we want to assess if the attribute values contained in the groups are similar.

2. Related Work

Exploratory visual analysis is a common approach for finding relationships and patterns in tabular data. Different established visualization techniques exist that support this goal, such as parallel coordinates [CvW11], parallel sets [KBH06], Table Lens [RC94], InfoZoom [SB00], Taggle [FGS*19], and StratomeX [LSS*12]—to name a few examples. Those techniques solely rely on users to visually evaluate the quality of the findings. In contrast, Voyager [WMA*16] recommends visualizations for the purpose of exploring tabular data. It does not provide any score for the associations between the attributes, but the breadth-oriented visual data exploration approach allows for combinations a user would not think of immediately. However, an additional tool like TourDino is needed to assess hypotheses based on visual observations.

A complimentary approach is to present scores and significance values to confirm visual patterns. This confirmatory analysis [KMSZ06] can be achieved by: (i) making use of scripting languages, (ii) embedding the scores inside visualizations used for exploring the data, or (iii) by providing a dedicated support view.

Scripting languages, like R and Python, allow users to calculate the scores for the different hypotheses. However, this requires scripting knowledge and training in statistics. **In-place embedding approaches** show the results of the statistical tests directly inside the visualizations that are used for the exploratory analysis. However, the space for embedding additional information is usually very limited and adding the information introduces additional visual clutter. SMARTexplore [BBS*18] uses a variation of the in-place technique by showing the significance of the patterns within the table. It uses a table-based visualization with a heat map color coding, where the items are grouped by a categorical attribute. The heat map shows the group's deviation from an attribute, e.g., by mean. SMARTexplore employs different similarity scores, depending on the data type, and is able to indicate the significant ones. In contrast to TourDino, it does not provide information about the statistical tests used or an additional explanatory visualization for the compared datasets. **Support views** present the information needed for confirming the visual patterns as an additional component, part of a multiple-coordinated view setup. The Rank-by-Feature framework [SS05], Guided StratomeX [SLG*14], and also TourDino are typical examples of this approach. Similar to TourDino, the Rank-by-Feature framework supports box plots and scatter plots as well as different scores. TourDino additionally provides the significance values for the scores. Even though the Rank-by-Feature framework provides different scores, they are applicable to numerical data only. In contrast, TourDino supports comparisons for the different combinations of categorical and numerical data. Each combination uses a suitable score, depending on the data. StratomeX [LSS*12] allows the visual comparison of different groups in a heterogeneous dataset and is able to find similarities using a query-wizard integrated in a comparative visualization [SLG*14]. TourDino also uses a wizard-like approach, but differs from Guided StratomeX in the sense that we only consider the displayed data and provide a more detailed description of the results.

The Visual Causality Analyst [WM16] enables users to analyze potential causal relationships by means of a node-link diagram. The attributes are represented as nodes and the scores are encoded on the edges. While both numerical and categorical attributes are supported, like in TourDino, the analysis of item sets is missing.

3. Pairwise Comparison Statistics

TourDino offers a comparison between multiple attributes and multiple groups of items. We separate these two fundamentally different tasks. An attribute comparison is calculated based on value pairs, and has therefore two equally sized value sets, e.g., tumor type and gender of a set of samples. This is different to comparing groups of items, where the value sets can have arbitrary sizes with independent values, for example, when comparing the tumor type of male and female samples.

We formulated a null hypothesis for each task we test. For the attribute comparison our null hypothesis is that the attributes describe distinct characteristics of an item, which are therefore dissimilar (i.e., independent). The second null hypothesis is that two groups of items are similar, as they are part of the same dataset that is described by the same attributes. As a result of our statistical tests, we report the score as well as the p-value, which reflects the probability that the null hypothesis is true. Results with a p-value

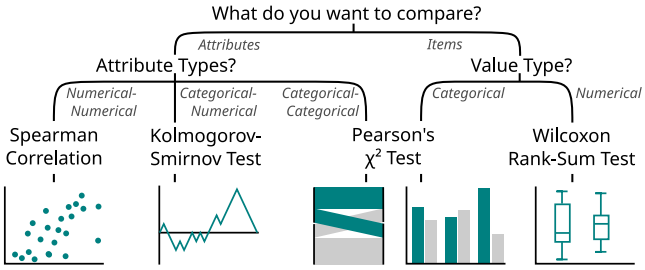

Figure 3: *Decision tree for choosing the statistical tests and the corresponding detail visualization for describing the pattern.*

below a predefined threshold—0.05 in our case—are assumed to be significantly different from our assumption. Hence, the lower the p-value, the more likely is an association between attributes or a difference between item groups, respectively. The statistical tests are chosen based on the user-selected task and the data type of the compared data subsets, as shown in Figure 3. In this work we focus on well established non-parametric statistical tests to create a general-purpose confirmatory analysis view that can be flexibly used in any tabular data visualization system.

To measure the strength of the association between two **numerical attributes**, we use the *Spearman Rank Correlation Coefficient* [Zar72]. We then transform the correlation coefficient to determine its significance with *Student's t-distribution* [Zar72].

To compare **categorical with numerical attributes**, we test if the items from any category are uniformly distributed across all items ranked by the numerical attribute. With a *Kolmogorov-Smirnov Test* [MLE*03] we find the maximum deviation of the category's ranked numerical values from a uniform value distribution. Additionally, we calculate the maximum deviation of 1,000 sets with randomized numerical attribute values. Whether the category has a significant effect on the ranks of the numerical values is determined by comparing its maximum deviation from uniform distribution with the deviations of the random sets.

For comparing **categorical data** we apply the *Pearson* χ^2 *Test* [Gin92, McH13]. When comparing categorical attributes, we conduct a *test for independence*, with the null hypothesis that there is no association between the attributes. The *test for homogeneity* is used when comparing categorical sets, with the null hypothesis that the categories are equally distributed in the sets. We use the χ^2 score to determine the association's significance and measure the strength with *Cramer's V* [Gin92, McH13], a normalized χ^2 score.

For comparing **numerical value sets**, we use the *Wilcoxon Rank-Sum Test* [Wil45] to determine if the two value sets stem from the same distribution. The scores of this test are approximately normally distributed, which we use to determine whether the sets differ significantly from each other [Wil45].

Conducting multiple of these tests leads to a higher probability of finding significant results by chance—known as the multiple testing problem. We inform the user about this risk by showing textual notes in the TourDino interface (see Figure 1b,e).

4. TourDino Support View

TourDino is designed as a support view for tabular data analysis allowing users to perform statistical tests on tabular data subsets, as demonstrated in Figure 1. The support view consists of three parts: (i) the *task selection* to choose the data subsets to be compared, (ii) the *significance matrix* summarizing the pairwise significance values of the tests performed, and (iii) the *detail visualization* presenting the details of a single pairwise combination.

Task Selection. Users can choose between two tasks to assess their findings from the visual exploration: *comparing attributes* or *comparing groups of items* (see Figure 2a). After selecting a task, users need to specify what data subsets should be compared. The attribute comparison takes two sets of attributes as input, where each attribute of the first set will be compared with every attribute of the second set. For the item comparison, users select which groups of items should be compared by selecting categories from the dataset. In addition to the items, users also select the attributes by which these items are to be compared. The growth type of brain cell lines, for example, can be compared with that of skin cell lines.

Significance Matrix. The significance matrix shows the probabilities with which the hypotheses of the applied pairwise statistical tests are true (see Figure 2b). We chose to display the probability values (p-values) instead of the scores. We argue that users can directly discard results that are not statistically significant. Further reasons are that the p-values are directly comparable between different statistical tests. We highlight matrix cells with a p-value below 0.05 by varying the brightness of the cell. The darker the background is, the higher is the significance. Additionally, results with a p-value above 0.1 are blanked out. The non-significant results are revealed when hovering over the respective matrix cell. A circle shown in the matrix cells indicates a comparison of a data subset with itself. We show a dash instead of the p-value if a statistical test is not applicable to a subset combination (see Figure 2b). This happens in two cases: (i) if a numerical attribute is compared with an attribute that has only one category, or (ii) if more than 90 percent of the compared data is invalid, e.g., null. We currently do not address the multiple testing problem in the significance matrix. While the results of multiple parallel tests could be corrected, sequential tests run by the user in multiple analysis sessions would still remain uncorrected and could ultimately lead to spurious results.

Detail Visualization. Users can select a matrix cell to get further information on a particular comparison (see Figure 2c). The information provided includes the name of the statistical test, a short description of how the test works, its score, the p-value, and a small detail visualization. The visualization is specific to each formulated hypothesis that we test and illustrates the similarities or differences that have been found (see Figure 3). For the Spearman Rank Correlation Coefficient, we depict the attributes' relationship using a scatter plot. We show the association between categorical and numerical attributes by plotting the deviation of the category's numerical values from a uniform value distribution. The shape of the curve indicates a potential enrichment. We visualize the association between categorical attributes using parallel sets. An independent attribute randomly splits each of its categories into the categories of a second attribute. The ribbons of the parallel sets show how the items are categorized in both groups. The distributions of categories in item groups are shown with relative frequency histograms. The χ^2 statistic rises with increasing difference between the histogram's bars of the respective sets. To show the differences between numerical value sets we use box plots. The more different the two sets are

in the box plot, the more likely it is that they do not stem from the same distribution.

5. Integration of TourDino into the Ordino System

We have integrated TourDino into Ordino [SGS*19], a web-based visual analysis tool for ranking and exploring genes, cell lines, and tissue samples. The heart of Ordino is an overview+detail tabular data visualization that supports comprehensive filter and aggregation capabilities. Our support view is part of the side panel, which provides various means to interact with the visualized data. We only compare the data subset that is currently visible in Ordino, such that the statistics are consistent with the visualization. To support users in switching back and forth between the main visualization and our support view, we highlight the compared data subsets in the main visualization. We therefore highlight the attributes and items that were compared, depending on the currently selected task, when hovering over a cell of the significance matrix. Figure 1 shows an example where two attributes that were compared with TourDino (b) are highlighted in the main table visualization of Ordino (g) by changing the background of the column headers to grey. In addition, when comparing groups of items, we highlight the attribute's values that are compared, and add a border to the categories that determine the two item groups.

6. Implementation

The TourDino support view is written as a client-side web-component in TypeScript and uses D3.js for creating the significance matrix and the detail visualizations. To enable a swift confirmatory analysis, we parallelize the calculations of the significance values and statistical scores using web workers and employ caching strategies to avoid redundant calculations. We provide the library (`https://github.com/Caleydo/tourdino/`) that contains the statistical tests and visualizations and the integration into Ordino (`https://github.com/datavisyn/tdp_core/`) as open source. The prototype implementation is publicly available at `https://tourdino.caleydoapp.org/`.

7. Case Study

This case study summarizes an analysis session carried out by a collaborator with a background in bioinformatics. While a part of the case study and its findings were originally reported in our previous work [FGS*19], we now demonstrate how TourDino helps to statistically confirm the validity of the results.

With the goal of identifying potential drug targets, the analyst conducts experiments with cancer cell lines, focusing on the important cancer genes *TP53* and *MDM2* in a subset of tumor types. Cancer cell lines are cultured cells that are derived from tumors that can proliferate indefinitely in the laboratory and are characterized by various properties, like tumor type and the set of genes that are mutated. *TP53* encodes the p53 protein, whose presence is known to suppress the uncontrolled division of cells. When *TP53* is mutated, it can lose its suppressing function, which results in tumor growth. Additionally, the inhibition of p53 through its interaction partner MDM2 can result in the loss of the suppressing function in cases where *TP53* is not mutated but *MDM2* is aberrantly highly expressed. The expression is a measure of the activity of genes.

First, the analyst wants to analyze if the expression level and the number of gene copies of *MDM2* correlate. An increased copy number of a gene can lead to a higher expression. The analyst loads the public Cancer Cell Line Encyclopedia (CCLE) dataset [BCS*12] into Ordino. Only a subset of tumor types is of interest, therefore the analyst filters for astrocytoma/glioblastoma (type of cancer in the brain), bone sarcoma, melanoma, and non-small-cell lung cancer (NSCLC). The analyst loads two attributes: the relative copy number and the gene expression of *MDM2*, and filters out the missing values. In order to check for correlation between these two attributes, the analyst opens the TourDino support view and performs an attribute-wise comparison. The analysis shows that the attributes are correlated (p-value < 0.001, compare Figure 1c), suggesting that in at least a subset of cell lines, a copy number alteration of the *MDM2* gene led to a change in expression.

As a next step the analyst wants to investigate if the expression and the copy number of *MDM2* correlate with the *TP53* mutation status. Hence, the *TP53* mutation status is loaded into Ordino, and cell lines with no mutation information are filtered out. Using TourDino, the analyst observes that the correlation between the *TP53* mutation status and the *MDM2* expression is significant whereas the correlation between the *TP53* mutation status and the *MDM2* copy number is not. This indicates that the actual gene expression is biologically more relevant than the higher gene copy number.

By inspecting the values of the significance matrix, the analyst also notices that the *TP53* mutation status correlates with the tumor type (p-value < 0.001) suggesting tumor type specific differences. In order to investigate this in more detail the analyst uses TourDino to compare the *MDM2* expression and *TP53* mutation status for all four tumor types (see Figure 1e). The analysis shows that, overall, melanoma cell lines have a significantly higher expression of *MDM2* and a lower *TP53* mutation rate compared to the other tumor types, especially NSCLC (see Figure 1f). This suggests different underlying mechanisms. However, the difference between, for instance, astrocytoma/glioblastoma and NSCLC is not significant.

8. Conclusion

In this paper, we presented TourDino, a support view for assessing visual patterns between attributes and item groups in tabular data that have been identified in an exploratory visualization. Attributes and groups of items can be compared with each other, using well established non-parametric statistical tests. The goal is to support non-experts in statistics, by providing clear information about the statistical significance of observed patterns in the data and by providing intuitive visualizations substantiating the statistical results.

In future work we plan to add more visualizations, e.g., to switch between a box plot and a violin plot. In addition we want to include parametric tests as they usually provide more accurate results in regards to the p-value, but require prior verification of their requirements. Currently we use TourDino to asses the significance between user defined data subsets, a possible third task could be added that provides the user with relationships of potential interest. Furthermore, we plan to correct the p-values for multiple testing.

9. Acknowledgments

This work was supported in part by Boehringer Ingelheim Regional Center Vienna, the State of Upper Austria (FFG 851460), and the Austrian Science Fund (FWF P27975-NBL).

References

[BBS*18] BLUMENSCHEIN M., BEHRISCH M., SCHMID S., BUTSCHER S., WAHL D. R., VILLINGER K., RENNER B., REITERER H., KEIM D. A.: SMARTexplore: Simplifying High-Dimensional Data Analysis through a Table-Based Visual Analytics Approach. *IEEE Conference on Visual Analytics Science and Technology (VAST) 2018* (2018). URL: http://kops.uni-konstanz.de/handle/123456789/43582. 2

[BCS*12] BARRETINA J., CAPONIGRO G., STRANSKY N., VENKATESAN K., MARGOLIN A. A., ET AL.: The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature 483*, 7391 (2012), 603–607. doi:10.1038/nature11003. 4

[CvW11] CLAESSEN J. H., VAN WIJK J. J.: Flexible Linked Axes for Multivariate Data Visualization. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '11) 17*, 12 (2011), 2310–2316. doi:10.1109/TVCG.2011.201. 2

[FGS*19] FURMANOVA K., GRATZL S., STITZ H., ZICHNER T., JARESOVA M., LEX A., STREIT M.: Taggle: Scalable Visualization of Tabular Data through Aggregation. *arXiv preprint* (2019). URL: https://arxiv.org/abs/1712.05944. 2, 4

[Gin92] GINGRICH P.: Association Between Variables. In *Introductory statistics for the social sciences*. Department of Sociology and Social Sciences, University of Regina, 1992, pp. 769–786. URL: http://uregina.ca/~gingrich/text.htm. 3

[KBH06] KOSARA R., BENDIX F., HAUSER H.: Parallel Sets: interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics 12*, 4 (2006), 558–568. doi:10.1109/TVCG.2006.76. 2

[KMSZ06] KEIM D. A., MANSMANN F., SCHNEIDEWIND J., ZIEGLER H.: Challenges in Visual Data Analysis. In *Conference on Information Visualisation (IV '06)* (2006), pp. 9–14. doi:10.1109/IV.2006.31. 2

[LSS*12] LEX A., STREIT M., SCHULZ H.-J., PARTL C., SCHMALSTIEG D., PARK P. J., GEHLENBORG N.: StratomeX: Visual Analysis of Large-Scale Heterogeneous Genomics Data for Cancer Subtype Characterization. *Computer Graphics Forum 31*, 3pt3 (2012), 1175–1184. doi:10.1111/j.1467-8659.2012.03110.x. 2

[McH13] MCHUGH M. L.: The chi-square test of independence. *Biochemia medica: Biochemia medica 23*, 2 (2013), 143–149. doi:10.11613/BM.2013.018. 3

[MF17] MATEJKA J., FITZMAURICE G.: Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017), ACM, pp. 1290–1294. doi:10.1145/3025453.3025912. 1

[MLE*03] MOOTHA V. K., LINDGREN C. M., ERIKSSON K.-F., SUBRAMANIAN A., SIHAG S., LEHAR J., PUIGSERVER P., CARLSSON E., RIDDERSTRÃËLE M., LAURILA E., HOUSTIS N., DALY M. J., PATTERSON N., MESIROV J. P., GOLUB T. R., TAMAYO P., SPIEGELMAN B., LANDER E. S., HIRSCHHORN J. N., ALTSHULER D., GROOP L. C.: PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics 34*, 3 (2003), 267. doi:10.1038/ng1180. 3

[Mun14] MUNZNER T.: *Visualization Analysis and Design*. CRC Press, Taylor & Francis Group, Boca Raton, 2014. URL: https://www.cs.ubc.ca/~tmm/vadbook/. 2

[RC94] RAO R., CARD S. K.: The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus + Context Visualization for Tabular Information. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (1994), 318–322. doi:10.1145/191666.191776. 2

[SB00] SPENKE M., BEILKEN C.: InfoZoom-Analysing Formula One racing results with an interactive data mining and visualisation tool. *WIT Transactions on Information and Communication Technologies 25* (2000), 455–464. doi:10.2495/DATA000441. 2

[SGS*19] STREIT M., GRATZL S., STITZ H., WERNITZNIG A., ZICHNER T., HASLINGER C.: Ordino: a visual cancer analysis tool for ranking and exploring genes, cell lines and tissue samples. *Bioinformatics* (2019). doi:10.1093/bioinformatics/btz009. 1, 4

[SLG*14] STREIT M., LEX A., GRATZL S., PARTL C., SCHMALSTIEG D., PFISTER H., PARK P. J., GEHLENBORG N.: Guided visual exploration of genomic stratifications in cancer. *Nature Methods 11*, 9 (2014), 884–885. doi:10.1038/nmeth.3088. 2

[SS05] SEO J., SHNEIDERMAN B.: A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization 4*, 2 (2005), 96–113. doi:10.1057/palgrave.ivs.9500091. 2

[Wil45] WILCOXON F.: Individual Comparisons by Ranking Methods. *Biometrics Bulletin 1*, 6 (1945), 80–83. doi:10.2307/3001968. 3

[WM16] WANG J., MUELLER K.: The visual causality analyst: An interactive interface for causal reasoning. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2016), 230–239. doi:10.1109/TVCG.2015.2467931. 2

[WMA*16] WONGSUPHASAWAT K., MORITZ D., ANAND A., MACKINLAY J., HOWE B., HEER J.: Voyager: Exploratory Analysis Via Faceted Browsing of Visualization Recommendations. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '15) 22*, 1 (2016), 649–658. doi:10.1109/TVCG.2015.2467191. 2

[Zar72] ZAR J. H.: Significance testing of the Spearman rank correlation coefficient. *Journal of the American Statistical Association 67*, 339 (1972), 578–580. doi:10.2307/2284441. 3