TaCo: Comparative Visualization of Large Tabular Data

Reem Hourieh, Holger Stitz, Nils Gehlenborg and Marc Streit

JOHANNES KEPLER UNIVERSITY LINZ

HARVARD MEDICAL SCHOOL

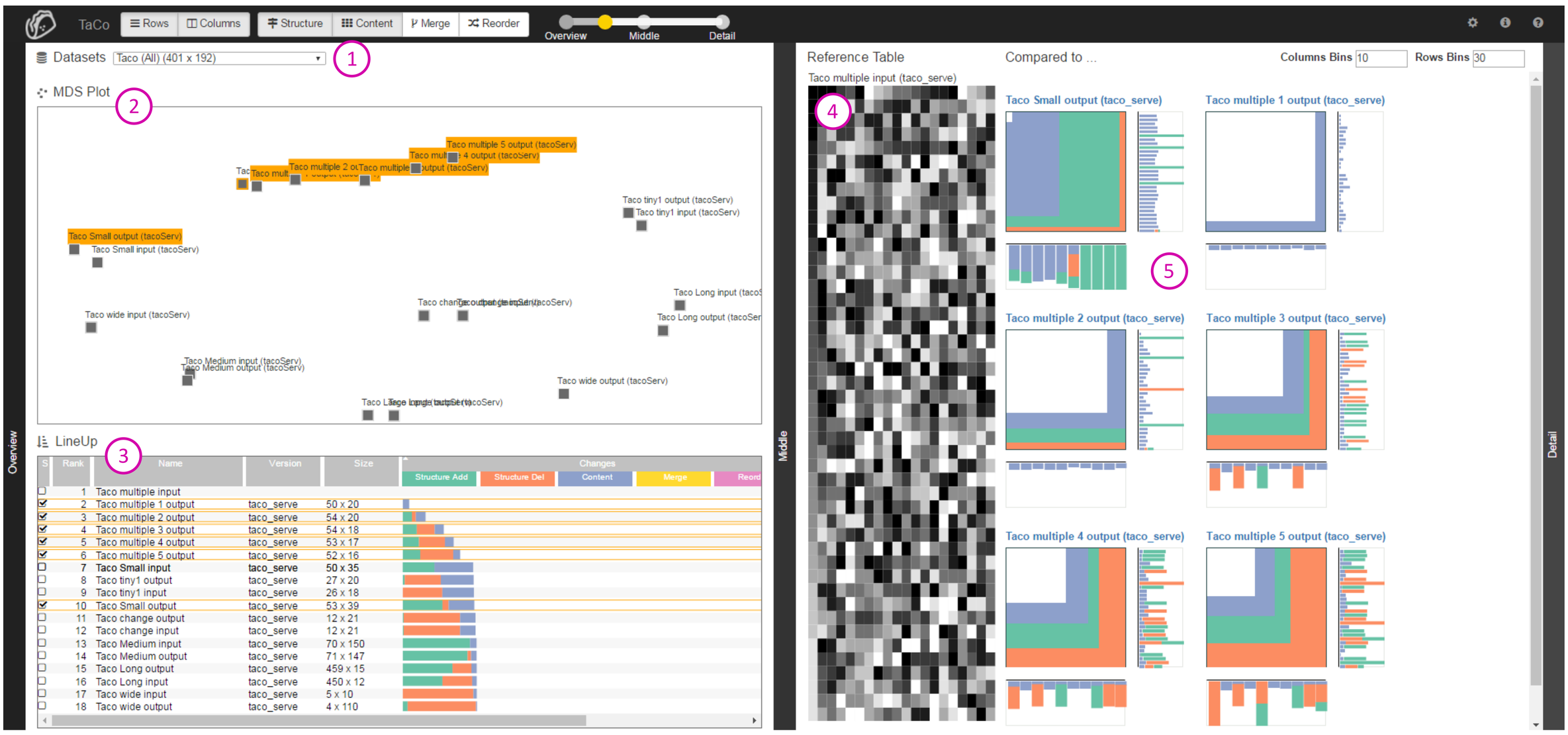

Figure 1: The multi-view interface of *TaCo* showing multiple instances of an artificially generated table. The overview on the left side (1) lets the user select a collection of tables that are plotted using (2) Multidimensional Scaling (MDS) based on the calculated similarity among the tables. (3) The user compares one selected reference table to all other tables in *LineUp*. The middle view on the right side shows (4) the reference table as a heatmap and (5) the aggregated differences to a selected group of tables for both row and column changes.

Tabular data plays a vital role in many different domains. In the course of a project, changes to the structure and content of tables can result in multiple instances of a table. *TaCo* (*Ta*ble *Co*mparison) is an interactive comparison tool that effectively visualizes the differences between multiple tables at various levels of granularity:

(1) Aggregated differences between all table instances (Fig. 1 (2))

(2) Differences between one table compared to all others (Fig. 1 (5))

(3) Detailed differences between two instances (Fig. 4 (2,3)).

Together with biomedical data analysts we elicited a series of tasks to be supported by an effective comparative visualization:

T1 Identify the type of changes as one of the four types

- *Structural changes* for added or removed rows/columns
- *Content changes* for modifications in a cell value
- *Reordering changes* for repositioned rows/columns
- *Merge changes* for combining multiple rows/columns together to yield only one row/column

T2 Compare two or more tables at various levels of detail (Fig. 2)

T3 Compare tables with regard to their dimensions (Fig. 3)

- *One dimensional histograms* encode the different changes for rows (Fig. 3 (3,5)) or columns (Fig. 3 (4,6))
- *2D-Ratio* visualization summarizes the changes in both directions (Fig. 3 (7))

Comparing large tabular data requires a two-part solution:

1. Calculating the difference between tables
2. Visualizing the difference in an effective and scalable way

Existing table comparison tools (e.g., DiffKit [1] and Daff [2]) generate a textual representation of the difference with basic color encoding, but do not scale to large tables. Furthermore, existing tabular comparative visualizations are usually task dependent, for example, for networks analysis [3] or database query comparison [4]. Other visualizations lack the ability to perform simultaneous row-wise, column-wise, and cell-wise comparison of tables [5,6].

[1] Panico J.: DiffKit. http://www.diffkit.org, 2016. Accessed: 2016-06-01
[2] Fitzpatrick P.: daff. http://paulfitz.github.io/daff/, 2016. Accessed: 2016-06-01
[3] Zhao J. *et al.*: *MatrixWave:* Visual Comparison of Event Sequence Data. In Proceedings of the ACM Conf. on Human Factors in Computing Systems (2015), ACM, pp. 259–268.
[4] Elmquvist N. *et al.*: *DataMeadow:* a visual canvas for analysis of large-scale multivariate data. Information Visualization 7, 1 (2008), 18–33.
[5] Lex A. *et al.*: Comparative Analysis of Multidimensional, Quantitative Data. IEEE Trans. on Visualization and Computer Graphics (InfoVis '10) 16, 6 (2010), 1027–1035.
[6] Behrisch M. *et al.*: Visual Analysis of Sets of Heterogeneous Matrices Using Projection-Based Distance Functions and Semantic Zoom. In Computer Graphics Forum (2014), vol. 33, Wiley Online Library, pp. 411–420.

Contact
holger.stitz@jku.at
caleydo.org
github.com/Caleydo/taco

Acknowledgements
This work was supported by:
- Austrian Science Fund (FWF P27975-NBL)
- State of Upper Austria (FFG 851460)

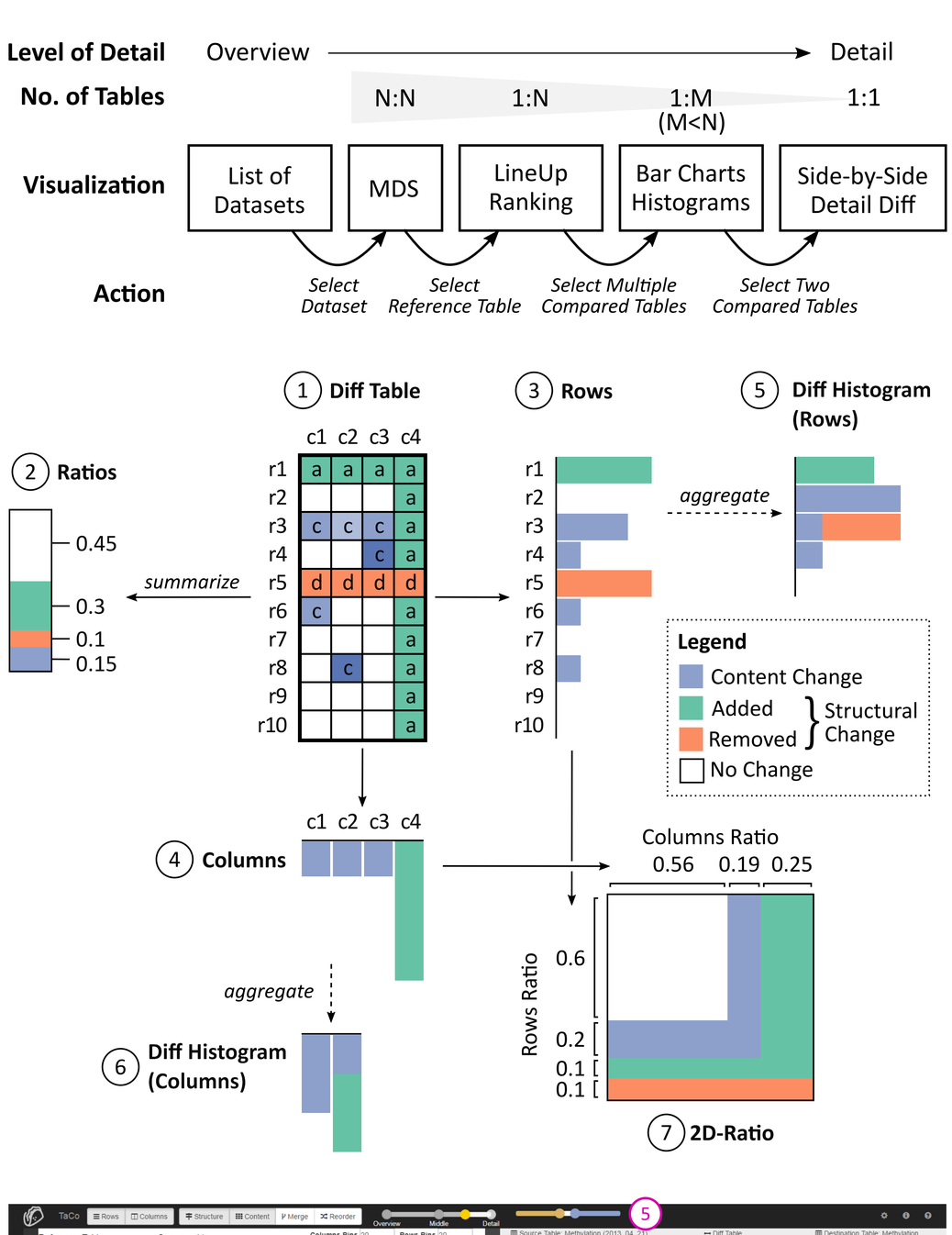

Figure 2
The four-stage visualization approach allows users to reduce the number of compared tables from one stage to the next stage, while increasing the details shown.

Figure 3
The difference between two tables is visualized as (1) a difference table. Changes are summarized on a per cell basis and visualized as a (2) ratios bar plot. The difference table can be aggregated for (3) row and (4) column directions separately. (5,6) Further aggregation for one direction is shown as a histogram. Summarizing changes for rows and columns results as a (7) 2D-ratio visualization.

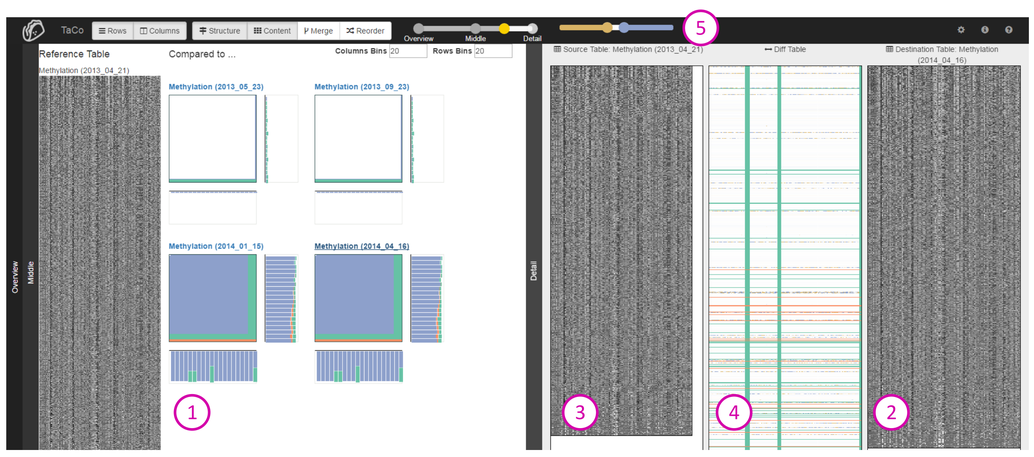

Figure 4
(1) The aggregation result of four TCGA tables is shown in the middle view. (2) One table was selected for a detail comparison and is visualized side-by-side with (3) the reference table as heat map and (4) the *diff table*. The color scale of content changes can be manipulated by the user with (5) the color-slider.