# SMARTexplore: Simplifying High-Dimensional Data Analysis through a Table-Based Visual Analytics Approach

Michael Blumenschein[1], Michael Behrisch[2], Stefanie Schmid[1], Simon Butscher[1], Deborah R. Wahl[1], Karoline Villinger[1], Britta Renner[1], Harald Reiterer[1], and Daniel A. Keim[1]

[1]University of Konstanz, Germany* [2]Harvard University, USA†

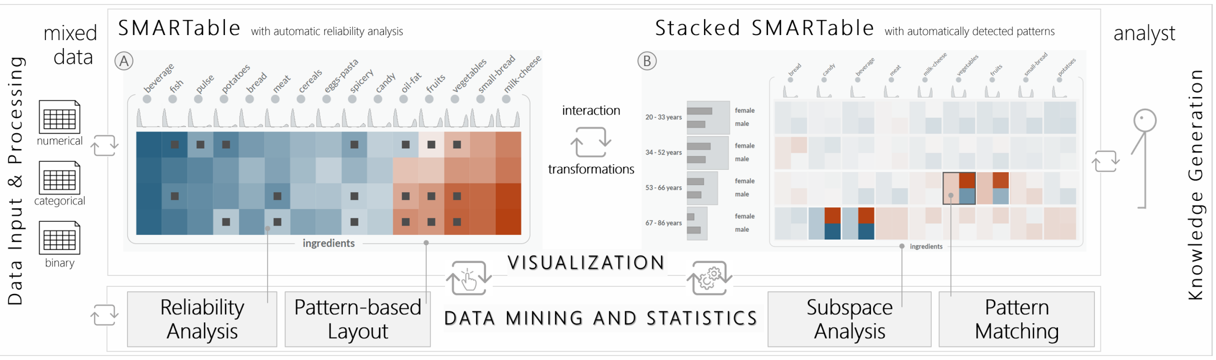


Figure 1: The visual representation of SMARTEXPLORE is a so-called SMARTABLE. Descriptors such as mean, variance, or deviation are computed, normalized per dimension or subspace, and mapped to a bi-polar or linear colormap. Manual and (semi-)automatic algorithms are executed through the visualization and support analysts in identifying and understanding clusters, correlations, outliers, and application-specific patterns in subspaces of the data. To increase trust in the patterns, statistical measures are computed on-the-fly and visualized along with missing values as visual overlays. Details on demand and a *stacked* SMARTABLE support detail analysis.

## Abstract

We present SMARTEXPLORE, a novel visual analytics technique that simplifies the identification and understanding of clusters, correlations, and complex patterns in high-dimensional data. The analysis is integrated into an interactive table-based visualization that maintains a consistent and familiar representation throughout the analysis. The visualization is tightly coupled with pattern matching, subspace analysis, reordering, and layout algorithms. To increase the analyst's trust in the revealed patterns, SMARTEXPLORE automatically selects and computes statistical measures based on dimension and data properties. While existing approaches to analyzing high-dimensional data (e.g., planar projections and Parallel coordinates) have proven effective, they typically have steep learning curves for non-visualization experts. Our evaluation, based on three expert case studies, confirms that non-visualization experts successfully reveal patterns in high-dimensional data when using SMARTEXPLORE.

**Keywords:** High-dimensional data, visual exploration, pattern-driven analysis, tabular visualization, subspace, aggregation.

## 1 Introduction

Users need to *find* and *understand* clusters, correlations, and complex patterns in high-dimensional (HD) data for many applications. Consider, for example, diabetes experts, seeking to understand the eating habits of individuals or groups of patients. Factors to explore could include similarities in meal ingredients between patients from different cultural backgrounds, whether location and environment influence the subjective enjoyment of a meal, or which combination of influences do (not) correlate with age. Often, the datasets are not only high-dimensional but contain a mixture of different data types, such as, numerical, categorical, and binary.

To analyze such patterns, the InfoVis community has acknowledged the need for visualizations and interactive tools to deal with the overwhelming complexity and the large amount of data. A broad number of approaches have been developed. However, they usually transform the data into abstract representations. Popular examples are Scatter plots, Parallel coordinates [29], and linear and non-linear projections, such as PCA [33] and MDS [60]. While these and other approaches have proven to be effective for the analysis of HD data, they often require long training for non-visualization experts and influence the analyst's trust in the revealed patterns [31]. Even after applying the concept of an abstract visualization, interacting with records and dimensions is seldom intuitive. Instead, it requires mental effort to interrelate records, dimensions, and values in the original format with the representation in the visualization and vice versa.

We present SMARTEXPLORE, an intuitive approach which injects visual analytics (VA) concepts into a table-based visualization. Rows represent records or record groups and columns, dimensions. A broad number of (statistical) measures, such as mean or deviation, can be computed, normalized, and represented with different colormaps, as shown in Fig. 1. Pattern analysis algorithms, reordering techniques, and interaction concepts support visualization experts and novice users, alike, to reveal patterns in large HD data. Whenever possible, algorithms are automatically applied to reduce the number of tedious or complex tasks. Our decision to develop an enhanced analysis system around a table representation is backed by the fact that HD data is usually given in a table format and that the majority of analysts are familiar with spreadsheet tools, like Excel. Over a long period, they have been trained to read such tables, modify, filter or reorder rows and columns; or compute new derivative measures, such as mean or variance. While table representations naturally have the disadvantage of an inflexible layout, recent tabular-based visualizations [17,22,23,64] have shown to be intuitive for a variety of user groups, even for complex analysis tasks. However, none of the existing approaches is designed to identify and understand patterns, such as clusters or correlations in HD (sub-)spaces.

The primary contribution of this paper is to *simplify the identification and understanding of HD patterns* through a table-based

*e-mail: first.last@uni-konstanz.de
†e-mail: last@g.harvard.edu

VA approach. First, we describe a set of 13 requirements for table-based visualizations supporting the identification and understanding of clusters, correlations, outliers, and complex patterns. Second, we introduce SMARTEXPLORE with the following four contributions:

**Automatic handling and aggregation of mixed data types.** SMARTEXPLORE supports datasets with a combination of numerical, categorical, and binary dimensions which are displayed in a consistent, unified representation. Hence, patterns across mixed types can be analyzed easily. Appropriate similarity functions, statistical tests, and algorithms are automatically selected and applied based on the dimension type and its properties such as the distribution.

**Simplification of complex data transformations.** SMARTEXPLORE implements complex data transformations such as (recursive) record grouping, pattern analysis, and subspace detection with a similar interaction design as known from classical table manipulations such as filtering and sorting.

**Automation of pattern identification and highlighting.** Based on visual template matching and (semi-)automatic table reordering, SMARTEXPLORE supports analysts to identify and understand patterns across a large set of dimensions and record groups.

**Trust-building through automatic reliability analysis.** To increase trust, SMARTEXPLORE automatically computes and visualizes uncertainty and statistical significance. An appropriate test is selected based on the dimension type, sample size, and distribution.

To guide the reader through the different visual mappings and various interaction techniques, we introduce a guiding dataset called `food`. The dataset contains 2,571 meals consumed by 99 participants over a period of eight days [66,67]. Each meal (data record) contains a combination of numerical, categorical, and binary dimensions: For example, the amount of kcal, sugar, vitamins (numerical), where and with whom the meal was consumed (categorical), and a binary representation of ingredients such as meat, fish, potato, and milk. Each participant occurs multiple times in the data with all of his/her consumed meals. Potential analysis questions for research include *"How age and gender affect the eating behavior of people?"* Due to data privacy restrictions, we removed dimensions with sensitive information for the examples in this paper. Although we use this dataset as a running example, SMARTEXPLORE can be applied to any HD dataset with homogeneous and mixed data types.

To evaluate the usefulness of our proposed technique, we implemented a prototype. The source code and a running version, which allows data uploads, is available on our website: `http://smartexplore.dbvis.de`. As a secondary system design contribution, SMARTEXPLORE stores the visualization properties *and* the applied interactions in the URL parameters of the web application. This URL allows easy sharing of findings and intermediate analysis results among researchers and fosters academic discussions. Many examples presented in this paper are marked with the icon ▦, which provides a hyperlink to our prototype with consistent settings. Hence, the reader can interactively explore the presented examples and continue the analysis from this point on.

Next, we collect requirements for table-based visualizations and discuss them in relation to related work in Section 3. In Section 4, we introduce the visual design of SMARTEXPLORE and the interpretability of visual patterns, user-guided analysis concepts (Section 5), and the fully automatic pattern matching and verification provided by SMARTEXPLORE (Section 6). Afterward, we present the expert case study evaluation and conclude the paper with a discussion.

## 2 REQUIREMENT ANALYSIS

SMARTEXPLORE has been developed in close collaboration with domain experts from the psychology domain. Although this is not the only analysis domain with HD datasets, psychologists are especially often confronted with large tabular datasets from user studies. Based on their common analysis tasks, we collected an initial list of requirements for tabular visualizations. To be of practical use to a broader number of domains, we generalized the requirements by our own experience and requirements by related table-based VA tools. We see our requirement analysis tailored towards the vision of a **pattern-driven analysis** of HD data, in which *finding* and *understanding* of clusters, correlations, and other patterns is of imminent importance. In contrast, Gratzl et al. [23] propose a set of ten requirements to compare rankings of data records, Perin et al. [49] specify eight requirements to encode, modify, and reorder raw data within a table, and the twelve requirements by Furmanova et al. [17] support the dynamic and hierarchical aggregation of rows.

Among all of these requirement lists, there is some overlap. All approaches call for a visual encoding of data values, manual or automatic rearranging and sorting of rows and columns, an interactive and responsive analysis refinement strategy, and data manipulation possibility. Most approaches require details-on-demand, applicability to datasets with missing values, and applying operations only on subsets of the data and/or dimensions.

However, while these and other requirements sound similar, their underlying purpose and implementation differs significantly (e.g., reordering to compare rankings vs. reordering to identify patterns like clusters). Therefore, we derived and generalized our requirements specifically for a pattern-driven analysis in HD data.

### General- and System Requirements

**R0: Support for Data- and Dimension Analysis.** A system should support *finding* and *understanding* the following basic patterns in the data- and the feature space: ***(a) clusters of data records*** according to the given feature space and a chosen similarity notion; ***(b) clusters of dimensions*** for a given grouping/clustering of data records; ***(c)*** linear and non-linear ***correlations*** among two or more dimensions; and ***(d) outliers*** in records, groups of records/clusters, and dimensions.

**R1: Persistent Representation.** To reflect the analysts' mental model and to mitigate potential misinterpretations, the visual representation of the data *and* the analysis results should be kept consistent.

**R2: Capabilities for Mixed Dimension Type Analysis.** To find *patterns across multiple mixed dimensions*, a system should be able to analyze *numerical*, *categorical*, and *binary dimensions* in a single view. Separate views for different data types avoid revealing relationships among those dimensions.

**R3: Interactive Response.** Interaction with a visualization should run smoothly. Whenever possible, results of user interactions should be shown directly and without large delays. For operations on records and dimensions (e.g., **R9**, **R10**) the user should be able to interact directly with the visible data (elements) instead of becoming lost in abstract or non-related handles.

### Scalability on Data Record- and Dimension Level

**R4: Grouping Data Records.** Users should be able to group a set of records into a group/cluster to reflect its similarity, and reduce the complexity of data. Besides manual grouping, established procedures, such as grouping by a given category, binning, and clustering of (a subset of) dimensions should be supported.

**R5: Value Aggregation.** All data records within groups/clusters should be meaningfully aggregated to support group comparisons. For every combination of group and dimension, multiple aggregated measures should be available. Users require standard statistical aggregations, such as *mean*, *median*, *min*, *max*, *variance*, and *standard deviation* for numerical dimensions. For all dimensions, users are typically interested in *distributions* plots.

**R6: (Visual) Encoding of Aggregated Values.** Aggregated values and distributions should be visually encoded, such that users can reliably assess their similarity and quickly retrieve relationships. The encoding should not only support a two-way comparison but also alleviate the challenging task of manifold comparisons (e.g., across multiple dimensions or clusters; see **R8**).

**R7: Grouping Dimensions into Subspaces.** A system should support users in finding subspaces that are semantically meaningful or revealing dimension and pattern relationships (**R0b**). Naturally, a dimension may be part of multiple subspaces. Thus, users should be able to interactively adjust subspace memberships (see **R10**) to reflect their personal understandings of the data.

#### Comparative Analysis of Record- and Dimension Level

**R8: Comparison of Records and Dimensions.** The visual arrangement, as well as the encoding (**R6**), should support users to compare records or record groups across large sets of dimensions. Simultaneously, a dimension or a subspace should be compared among multiple record (groups). The concurrent comparison of records and dimensions supports the user in comprehending the visible patterns.

#### Data Handling- and Transformation

**R9: Operations on Record Groups.** Users should be able to operate intuitively on record groups to find and understand patterns (**R0**), thereby facilitating the comparison of records and dimensions (**R8**). The following group operations are required: ***(a) select*** and ***highlight***, ***(b) filter*** and ***remove***, ***(c) change ordering*** (manual) and ***automatic sorting*** based on similarity or by dimension/subspace, ***(d) merge*** one or more groups, and ***(e)*** extend grouping by ***recursively grouping*** records within a cluster.

**R10: Operations on Subspaces.** Users should be able to interact with dimensions and subspaces to facilitate records and dimensions comparisons (**R8**). All operations should be provided for each individual dimension and subspace: ***(a) select*** and ***highlight***, ***(b) remove***, ***(c)*** change ***ordering*** of subspaces and dimensions within a subspace and ***automatic sorting*** based on similarity, ***(d) add new*** subspaces, and ***(e) copy and move***, dimensions across subspaces.

#### Reliability and Trust

**R11: Reliability of Perceived Patterns.** Users require support to assess the reliability of findings. In particular, users need visual/algorithmic support for assessing: ***(a) missing data***, ***(b) too small groups***, or ***(c) statistically (non-)significant*** patterns. A system should be able to remove record groups or dimensions, classified as unreliable by the user (see **R9** and **R10**).

**R12: Provenance of Visualizations and Interactions.** A system should support storing intermediate analysis results and their associated visualizations, including all applied operations. Hence, results can be shared among researchers and analysts can reiterate previous results, or follow promising new analysis paths (see also **R11**).

## 3 Related Work

In the following, we delineate SMARTEXPLORE from other tabular-based and general HD visualization approaches, and show similarities to existing works for mixed datasets and trust-building in VA.

### 3.1 Table-based Visualizations

The most commonly used representation for HD data is a spreadsheet, such as Microsoft Excel [20] or Google Sheets [19]. These tools typically allow a wide range of row and column interactions, and let the user augment current analysis results with basic visualizations, e.g., bar charts and scatter plots. More interactive approaches support a larger set of visualizations, e.g., Tableau [58], Spotfire [57], Power BI [44], and JMP [52]. All these tools use tables for their data representation and use more or less intuitive mappings into different visual representations. Tableau and Spotfire focus on visual analysis, while JMP represents the model-building and statistical analysis spectrum. However, the approaches miss a tight integration between algorithmic support and visualization. Although the sophisticated visualizations for parts of the analysis process are usually linked to the table, they still require frequent mental model adoptions and changes. Table lens [50] is one of the first approaches to overcome this problem. The data stays in a table format, but the values in the rows and columns are approximated by sparklines [61]. An interactive focus+context approach enlarges rows and columns of interest. FOCUS [56] extends the idea through interactive queries that focus on data areas of interest.

The three approaches most related to ours are Bertifier [49], Taggle [17,18], and $I^F$, $F^I$-Tables [64]. Bertifier implements the idea of Bertin's reorderable (glyph) matrix [6] into an interactive tool. As in the original work, row and column ordering is the primary interaction concept for identifying patterns. Yet, it does not allow aggregating records or dimensions. Taggle features hierarchical aggregation of records, but compared to SMARTEXPLORE, the analysis goal differs. Taggle is used to compare aggregations on different granularity levels, rather than finding patterns across a large set of dimensions. $I^F$, $F^I$-Table uses two interlinked tables to compare records across a large set of dimensions and vice versa.

The visual representation of SMARTEXPLORE is also inspired by recent work in matrix visualization [1,14,36,62]. Most matrix visualizations are static and cannot be interactively adjusted. In particularly, matrices mostly feature algorithmic approaches that optimize the layout for one particular visual pattern. However, as also stated in a recent survey [4], these visual patterns do not necessarily align with the user's analysis question. As envisioned in [4], SMARTEXPLORE implements a more adaptive, user-guided process, which goes further than just drag&drop approaches, such as presented in [54]. Many other table-based visualizations exist. However, their core analysis tasks and contributions differ from SMARTEXPLORE. LineUp [23] identifies multi-attribute rankings in a table-like representation containing stacked bar-chart-like visualizations. Similarly, Podium [68] lets the user adjust rankings update the weights of the underlying ranking function. Taco [46] visualizes change over time within an aggregated table. A popular technique to visualize the relation between sets is UpSet [37]. Domino [22] lets users interactively combine, arrange, and extract subsets of data from different sources within a combined table-based view.

The last category of related table-based approaches are tools that let the user find patterns using sophisticated sorting algorithms. Examples are SimulSort [28], Matchmaker [38], and StratomeX [39].

In recent years, the InfoVis community presented a multitude of novel table-based visualization and VA systems. Most of these systems show that a representation in a table supports the users in the analysis process. However, the focus of the presented techniques is different from SMARTEXPLORE, as it combines sophisticated aggregation and grouping features, with pattern matching and an automatic reliability analysis.

### 3.2 Visualizations for High-dimensional Data

The community has presented many approaches for the analysis and visualization of HD data. Each approach has its advantages and disadvantages, and their discussions fill entire surveys [40].

With respect to our work, most approaches present specific solutions and trade-offs by focusing either on data vs. visual scalability, and complexity vs. understandability. For example, the seminal work of Inselberg on Parallel coordinate (PC) [29] advanced the field by focusing likewise on dimension and record scalability. Many improvements for PC have been proposed. For example, highlighting density [43,72] and quality metrics [5] which reduce visual clutter [13,48], or reveal specific patterns [12] by reordering the axis. Analogue to the idea of Ankerst et al. [3], the dimensions of SMARTEXPLORE can be reordered by visual similarity or particular visual patterns across multiple dimensions.

Orthogonal projections, such as in its bivariate form in Scatter plots, are also used for HD analysis. Here, the dimension interpretability and scalability is sacrificed for a better understandability of data record relations. Yet, a large set of possible dimension combinations has to be assessed for its usefulness. Tatu et al. and Albuquerque et al. present image and data-space quality metrics to quantify patterns in large sets of Scatter plots [2,59]. Non-linear [60,65] and linear projections [33] are classic approaches for HD analysis and

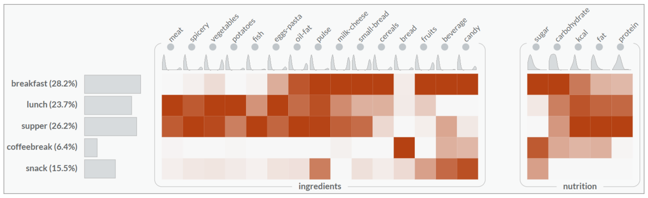


Figure 2: Example of a SMARTABLE: meals are grouped by type (rows). A color gradient (white → red) is used to describe how often a specific ingredient and nutrient (dimension, column) is part of a meal.

visualization. In the context of dynamic graph analysis, van den Elzen et al. use a 2D projection (t-SNE) of topologically similar graph snapshots for their argumentations [63]. Visual complex glyph designs layouted with 2D projections are presented e.g., in [9, 35].

To improve the understandability of HD datasets, navigation and user-guided exploration techniques have been presented recently with LDSScanner [70], in Subspace Voyager [69], and in Dimension Projection Matrix/Tree [71]. Fernstad et al. [15] propose a quality-metric guided framework for exploratory dimensionality reduction. Based on a large set of quality metrics, users can interactively rank and weight variables to reveal HD patterns. SMARTEXPLORE also supports metrics to identify visual patterns in the aggregated table. However, the metrics are computed in the image space which mimic the perception of analysts [5].

Only few approaches tackle HD dataset with mixed data types. The reason for this is the incompatibility of types w.r.t. distance functions and visual mappings. Often, different representations for different types (e.g., [17]) are used. Approaches focusing on the relation between data records typically apply the Gower distance [21] and project the data using MDS [11] into a Scatter plot. To identify relations between dimensions, categories are transformed into comparable numbers based on an application-dependent ordering or distribution [24]. The transformation can be done automatically [51] or with the help of analysts [32]. SMARTEXPLORE also transforms the distribution of categories into a numerical representation and visualizes it with the same encoding as numerical and binary dimensions within the table. This allows an easy identification of outliers and patterns in record groups across large sets of mixed dimensions.

Building trust in analysis results requires showing potential uncertainty along the analysis process. SMARTEXPLORE presents an automatic reliability analysis, which *automatically* determines and executes the correct statistical test from a set of 15 mathematical models. SMARTEXPLORE is influenced, by Correa et al.'s work on reflecting uncertainty aspects with visual mappings [10]. Similarly to Buchmüller et al. [7], we use a semi-transparent random noise and colored overlays to represent the uncertainty of the computed descriptors.

## 4 VISUAL DESIGN IN SMARTEXPLORE

We introduce the SMARTEXPLORE technique and show how it addresses the specified requirements. In Section 5 and 6 we show how interaction and automatic algorithms can support users when finding and exploring high-dimensional patterns.

### 4.1 Visual Design for Aggregated Features

We define the basic visual representation of SMARTEXPLORE as SMARTABLE: data records can be grouped into clusters and dimensions to meaningful subspaces. The values in every record group are aggregated to its distribution or (statistical) measures such as mean or variance. We show an example of a SMARTABLE in Fig. 2 based on the `food` dataset. The analyst has grouped the meals by type. The first row contains breakfast meals, then lunch, supper, meals consumed during coffee breaks, and snacks. Only dimensions in the ingredient and nutrition subspace are visible. The color gradient (white → red) describes the average number of meals containing a particular ingredient. Analysts can clearly see that ingredients towards the right dominate breakfast meals (except for the dimension *bread*), and ingredients on the left are mostly consumed during lunch and supper. On the left side of the table, users can compare the size of the record groups with the help of a histogram. The distribution of values in every dimension is visualized as distribution plot on top of each dimension. During the entire analysis, SMARTEXPLORE remains in this representation (**R1**) to reflect the mental model of users. Different visual overlays help to visualize mixed dimensions in a homogeneous view (**R2**) and highlight the results of automatic algorithms for pattern reliability analysis (**R11**).

We do not claim any superiority of our approach compared to established HD visualizations. However, in this paper, we show that a table contrasts well with abstract visualizations when equipped with VA tools. The well-known structure of rows and columns corresponding to records and dimensions respectively, appears advantageous for visualization experts and non-experts alike. It supports analysts in easily understanding the visual structures, and intuitively operate on record groups (**R4**, **R9**) and dimensions (**R7**, **R10**). Although the layout of tables is restricted to a grid, table cells can be arbitrarily complex. We show this aspect with our automatic reliability glyph (**R11**), which descriptively summarizes statistical reliability tests.

#### Data Record Grouping

Record grouping and clustering (**R4**) are useful means for spotting global patterns in the data and reducing the complexity. Additionally, analysts are often interested in understanding the properties of a given natural grouping in the data, e.g., compare different meal types as shown in Fig. 2. SMARTEXPLORE supports different record grouping strategies, useful for different applications and data types.

**Existing groups.** Categories naturally provide semantic groupings. All records with the same category can be combined into one group.

**Binning.** To find groups in numerical dimensions, *equal-width* or *equal-height binning* can be applied. In our implementation, we show an interactive preview in which the user can freely experiment with the bin size and instantly see the binning result in a histogram.

**Clustering.** An algorithmic solution for finding groups of data records is to apply clustering. However, not all dimensions in a dataset may be relevant to determining application dependent clusters. Therefore, a user can select a subspace of dimensions to be considered. Finding a good parameter setting and a good number of clusters, in particular, is challenging. In SMARTEXPLORE, we compute a hierarchical clustering [25] and let the user interactively adapt cluster numbers using a threshold in a visualized dendrogram. For numerical subspaces, a Euclidean distance, for subspaces with mixed data types (**R2**), the Gower metric [21] is used.

#### Descriptors: Aggregated Values of Record Groups

Every record or record group is represented by one row in the SMARTABLE. Comparing record groups across dimensions, and dimensions across record groups (**R8**) is a central analysis task for SMARTEXPLORE. Consequently, all values within a group need to be *aggregated* (**R5**) and visually *encoded* (**R6**) to foster comparability. We define *aggregated values* synonymously as *descriptors*.

In our prototype, we decided to implement the following descriptors: For **numerical dimensions**, we support the mean, median, min, max, variance, and standard deviation. The values in **binary dimensions** are $true = 1$ and $false = 0$. As a descriptor, we compute the mean, which corresponds to the percentage of records with the value *true*. This descriptor is also used in the example in Fig. 2 to show the frequency of ingredients for meal types. In **categorical dimensions**, a user is oftentimes interested in the distribution of categories. Here SMARTEXPLORE supports the visualization of the distribution as an overlay.

**Descriptors for mixed dimension views.** The aim of SMARTEXPLORE is to visualize all dimensions, independent of its type in a consistent representation (**R2**). Therefore, the aggregated values of mixed dimensions need to be represented by a descriptor that can

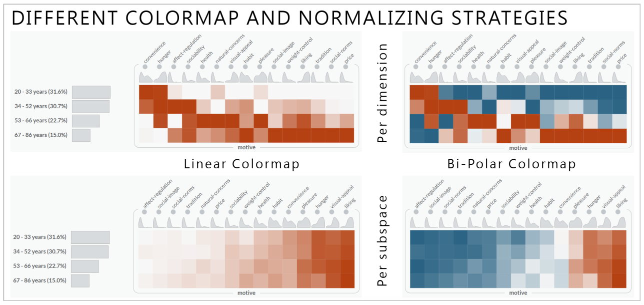


Figure 3: Comparison of *linear* (left) and *bi-polar* colormap (right) with *normalizing per dimension* (top) and *per subspace* (bottom). Visualized *descriptor*: *mean* per record group. All four examples represent the same data. The ordering is based on visual similarity.

be compared across data types. SMARTEXPLORE proclaims a so-called *deviation descriptor*. It measures the deviation between the descriptor of a record group and the same descriptor for the entire dimension. For example, for numerical and binary dimensions, we compute the difference between the mean of a group and the mean of the dimension. The deviation descriptor of categorical dimension is defined as the Euclidean distance between the frequency histogram of a dimension and within one group. While the deviation is computed differently for all data types, the intuitive understanding is the same: the value or distribution differs (much) from the overall dimensions. Hence, users can quickly spot dimensions or groups with less/more deviation and then continue the analysis with different descriptors. An example can be found in for the meal subspace.

### Visual Representation of Descriptors

To allow a fast comparison between record groups and dimensions, SMARTEXPLORE encodes (**R6**) the computed descriptors by color (**R8**). Similar measures are represented by similar colors, thus helping analysts to spot patterns. In the literature, there is a myriad of guidelines which help users to select an appropriate colormap for a specific task (e.g., [8, 45, 73]). During the development of SMARTEXPLORE, as well as many discussions with potential and active users, we found that there are two classes of colormaps that appear to be useful for analysis: *linear* and *bi-polar* as depicted in Fig. 3. We implement a linear colormap white (low) → red (high), and a bi-polar colormap blue (low) → white → red (high value). While linear colormaps enable users to directly compare two descriptors, bi-polar colormap are a great tool for identifying descriptors with high and low values.

### Normalizing Strategies for Descriptors

Normalizing is essential for promoting the visual prominence of patterns in SMARTEXPLORE. Since we apply the concept of aggregations with respect to records and dimensions, we need a flexible mechanism to normalize distributions. The intuition of the two implemented strategies is given in Fig. 3. Per default, descriptors are **normalized per dimension** , considering the descriptors of all record groups *within one dimension*. This strategy supports users to easily spot high, middle, or low values, but sacrifice the descriptors' comparability across dimensions. As a result, users can find patterns across multiple dimensions - even of dimensions with a different scale. Users can directly compare descriptors by **normalizing across dimensions** of a subspace (Fig. 3 bottom). In this case, the min and max within an entire subspace are used. This strategy only makes sense if all dimensions have semantic connections and have the same dimension scale. However, if this strong requirement holds, we allow users to derive conclusions from this fact, e.g., quantify descriptors across multiple dimensions. While scaling can be checked automatically, semantic interpretability needs to be determined by the user. Descriptors can be normalized *linearly* or *logarithmically*. Additionally, we allow users to inject domain knowledge by manually setting min and max; e.g., for a manual outlier correction or scale capping.

### Subspaces and Dimension Grouping

SMARTEXPLORE allows to group a subset of dimensions into a so-called *subspace* (**R7**). Every subspace contains at least one dimension and has a label which can be set by the user. A particular dimension can be part of more than one subspace. The reason for grouping dimensions into subspaces is twofold: First, it reduces the complexity of the visualization by introducing *visual gaps* between groups of dimensions that are semantically meaningful. Second, all visualization properties, such as normalizing strategy, colormap, sorting of dimensions etc., can be adjusted per subspace. This means, a user can group dimensions that should be treated similarly into a subspace, and select different properties for different subspaces.

## 4.2 Visual Design for Stacked Record Grouping

So far we have considered the *elementary aggregations* of the data records: A single dimension or a clustering algorithm determines the grouping of the data records. Every aggregated group is visualized as one row in the SMARTABLE. In many applications, users are interested in details of the aggregated rows. Consider for example Fig. 1 (B) . Records are grouped by *age* into four groups. Users may now be interested in *similarities/differences* between male/female *within each aggregation*. To support this analysis task, SMARTEXPLORE implements *stacked aggregations* (**R4b**). Each age group is further aggregated into a second level by the dimension *sex*. The distribution of both aggregation levels is shown by the histograms on the left side. The descriptor (here: mean) of the first aggregation level is represented by the upright rectangular and the descriptors for the stacked aggregation by the smaller squares on the right side. The stacked aggregation help users to analyze whether there is a difference in the descriptor when considering a more fine-grained aggregation. In Fig. 1 (B), we can see that the mean value of the dimension *vegetables* for the age group 53-66 years (marked) is light red. Male participants within this group have a much smaller mean value (dark blue) compared to female participants (dark red).

Stacked aggregations can also be created for more than two groups in the second level. For example, we can aggregate the records first by the attribute *age* into four groups and then by the *meal type* into five groups in the second level .

## 4.3 Interpretation of Patterns in SMARTExplore

Every system with an elaborated visual design allows identifying and describing *how* the occurring visual patterns need to be interpreted and how they support the analysis process. An overview of SMARTEXPLORE's most common visual patterns can be found in Fig. 4 and 5, along with a mapping of the analysis task (**R0**).

### Patterns Within and Across Dimensions

We have to distinguish between patterns existing *within a single dimension* and patterns *across multiple dimensions*. *Within dimensions* refers to patterns within a single dimension based on the current record grouping. For example, we see correlations, clusters, and outliers for dimension $B$ in Fig. 4 (left). Patterns *across dimensions* allow relating and comparing descriptors across multiple dimensions - typically all dimensions of a subspace. For example, we can see correlations, clusters, and outliers for different record groups across the dimensions $B, \ldots, K$ in Fig. 4 (right).

### Understanding Correlations

Analysts have to distinguish two notions of correlations (**R0c**):

**Correlations between dimensions and record groups** stand out as color gradients within one dimension (e.g., dimension $B$). Assuming that aggregated rows are in ascending order, Fig. 4 (a) shows dimensions with *positive*, *negative*, and *non-linear correlations*. Dimension groups can be clustered into a subspace to foster interpretability (e.g., dimensions $C$, $D$, and $E$).

**Correlations across dimensions** are independent of an ordering of the aggregated rows; however, they require an ordering among

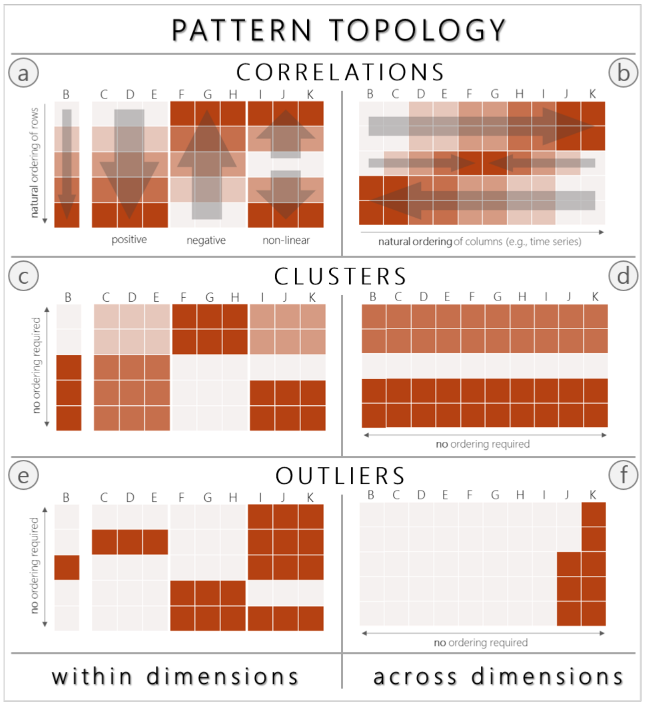


Figure 4: Most important patterns in SMARTABLE. Correlations, clusters, and outliers can occur *within* and *across* dimensions.

the dimensions (e.g., dimensions representing the values of a time series). Fig. 4 (b) shows an example of correlations across the dimensions $B, \ldots, K$.

### Understanding Clusters

Three cluster types can be analyzed with SMARTEXPLORE:

**Clusters of similar dimensions.** Visually similar dimensions can be clustered into a subspace (**R0b**). Hereby, the structure of the pattern does not matter. In Fig. 4 (a) and (e), dimensions with the same correlation and outlier pattern are clustered (e.g., $F$, $G$, and $H$).

**Clusters within a dimension.** Data records or record groups with similar descriptors can be perceived as a cluster (**R0a**). For example, the same value distribution in dimension $B$ is shared among the first two and last three record groups in Fig. 4 (c).

**Clusters across dimensions.** Fig. 4 (d) depicts three clusters of record groups which are described by all dimensions of the subspace.

### Understanding Outliers

An outlier is defined as a computed descriptor which differs substantially from all other descriptors. Based on the normalizing strategy, all descriptors of a dimension, or the descriptors of all dimensions of a subspaces, need to be taken into consideration when determining an outlier. Two types of outliers can be analyzed:

**Outliers within a dimension**. Fig. 4 (e) depicts an outlier in the dimension $B$ in the third record group.

**Outliers across multiple dimension** can be found in Fig. 4 (f). To find this pattern, dimensions need to be normalized per subspace. All record groups of a dimension can be considered outliers (dim. $K$), but also only as a subset of the groups, as shown in dimension $J$.

### Understanding Patterns in Stacked Aggregations

Stacked aggregations help users retrieving commonalities and differences across subcategories. Generally, there are two possibilities: stacked and base descriptors have the *same color*, or have a *different color*, implying descriptor similarity or dissimilarity, respectively. As shown in Fig. 5, correlation, outlier, and cluster patterns exist in stacked groupings. Of course, the pattern depends on the ordering of the records in the stacked aggregation.

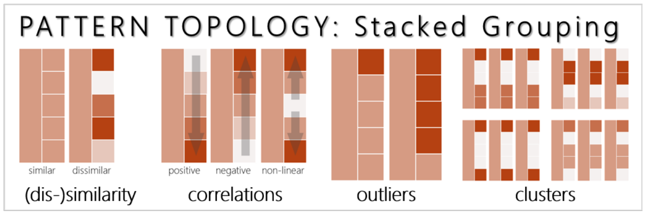


Figure 5: Common patterns in stacked SMARTABLE. Most importantly is whether the base and stacked descriptor are similar or not. If they are dissimilar, correlations, outliers, and clusters can be present.

### Application Specific Patterns

In our experiments with psychologists, we came across interesting patterns that cannot universally be described by a topology, instead, they depend on the analysis. For example, multi-modal distributions within dimensions, or 'expected' outliers in clusters of records or dimensions. SMARTEXPLORE can also be used to find and understand such application dependent patterns. However, analysts are necessary for the pattern interpretation.

## 5 USER-GUIDED ANALYSIS IN SMARTEXPLORE

SMARTEXPLORE provides easy to use interaction concepts to find interesting patterns, analyze records across dimensions, and dimensions across records, which are introduced in the following.

### 5.1 Interaction on Record Groups (R9)

As known from spreadsheet applications, users can *select and highlight* one or more record groups to compare the descriptors across all dimensions. By means of drag&drop, the *order of rows* can manually be *changed*, e.g., to compare record groups temporarily. Users can change the ordering of clusters to reflect semantic relationships, such as the temporal order of meals (Fig. 2). After the grouping, analysts can also select a dimension and reorder the groups based on the dimension's descriptors. This 'sorting operation' helps identify potential correlations, clusters, and outliers of the selected dimension.

Users can *delete* entire record *groups* to remove outliers, and to tailor the analysis to a specific task. Record groups can be *merged*, the grouping granularity can be *customized*, and non-uniform binning of records is supported. One application scenario is shown in Figures 1 and 3, where participants were grouped into ten bins, and then manually merged into four application-specific groups. During the analysis, users can freely adapt the grouping and *apply stacking*.

### 5.2 Interaction on Dimensions and Subspaces (R10)

Grouping dimensions to subspaces provides a useful mechanism to reduce the complexity of a HD dataset. While we still keep all dimensions for our analysis, we introduce a *visual* gap that separates subspaces. Thus, we are subdividing the SMARTABLE into small, semantically meaningful, and cognitively graspable subsets.

Analog to record groups, both dimensions and subspaces can be *selected*, *highlighted*, and *rearranged* for a better comparison. Assuming the user has built a mental model of the underlying relationship, SMARTEXPLORE allows *dragging&dropping* dimensions into interpretable subspaces, as shown in Fig. 2.

The *ordering* of subspaces along the x-axis can be changed by drag&drop. This enables the user to arrange subspaces close together, fosters comparability and understandability, or to drag subspaces to prominent positions at the start or end of the table. Subspace can be *deleted* (irrelevant for analysis) or *cloned* (show in other context), and *new* subspaces with a customized name can be created on-the-fly (further semantic relationship). Users can *copy* or *move* dimensions from one subspace to another in order to reflect their interrelation in the current analysis task. The properties of a dimension (e.g., colormap, computed descriptor, normalizing strategy) can either be changed globally for all dimensions, or per subspace. For example, a user can clone a subspace, and visualize its different statistical facets, e.g., its mean and variance. Within a subspace users reorder dimensions by dragging&dropping or *removing* them entirely.

A semi-automatic grouping of dimensions based on their similarity helps with deriving non-obvious subspaces. We apply a *hierarchical clustering* on all dimensions and map similar dimensions to a subspace. As in record grouping, a slider allows interactively changing the *granularity* of clusters. SmartExplore applies a Euclidean distance between all descriptors of two dimensions. Intuitively, this means visually similar dimensions will end up in a cluster. The Euclidean distance can be weighted by statistical significance.

As one algorithmic contribution, SmartExplore supports *semi-automatic pattern highlighting* and *sorting*. Users can select a dimension of interest and sort the remaining dimensions based on similarity. Our pattern matching algorithm can also highlights all dimensions similar to this selection. Finding similar dimensions can help users to identify patterns across multiple dimensions. The similarity search can be restricted to a subspace or can be applied on all dimensions of the dataset. Similar to the hierarchical clustering of dimensions, a Euclidean distance, optionally weighted by significance, is used to determine the similarity between two dimensions. SmartExplore proposes the number of dimensions to be highlighted based on the calculated distance distribution. The user can modify the expected highlighting accuracy (precision vs. recall) with the help of a slider. Highlighted dimensions can be *copied* or *moved* to a new or different subspace. This analytic guidance feature allows users to define subspaces with specific visual patterns.

### 5.3 Interaction on Aggregated Descriptors

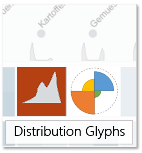
Semi-automatic pattern highlighting is also implemented for descriptors. In SmarTable, users can select a *cell of interest* and highlight the k-nearest neighbors. Searching for similar descriptors is most useful for stacked aggregations, as shown in Fig. 1 (B). For the pattern highlighting of stacked descriptors, users can either consider only the patterns in the stacking, and ignore the value of the base descriptor (as applied in Fig. 1 (B)), or choose a $50:50$ weighting to incorporate the base and the stacked descriptors. Both options have valid argumentations based on their use case. Here also, a slider allows defining the degree of (dis-)similarity, which should be considered in the analysis.

### 5.4 Details on Demand for Record-Level Analysis

A computed descriptor represents a data distribution in one aggregated value. However, the entire distribution should often be taken into account to obtain a valid pattern interpretation.

#### Distribution Overlay

The user can add a distribution overlay on top of each visualized descriptor ▦. A kernel-density estimation is used for numerical dimensions, a histogram for categorical and binary dimensions. The kernel-density curve depends on the parameter *bandwidth*. We estimate a good selection of the method proposed by Silverman [55]. Additionally, the user can change the kernel-density curve with a histogram, and, for the categorical dimension, change the histogram into a glyph representation, which is inspired by Star Glyphs [53]. The overall distribution of a dimension gives users a first impression of the data and can help interpreting measures and removing outliers.



#### Table Lens and Tooltip

Often, a user is interested in seeing all distribution details for one record group and/or one dimension. For this purpose, SmartExplore implements a tooltip for a single cell and a table lens [50] for entire rows/columns. Hovering over a cell depicts the data distribution for the overall dimension and the record group, along with information about missing values, and results of statistical tests, such as the $p-value$ and the applied test (see Section 6.3 for details). Hovering over a record group or dimension enlarges the visualized descriptors and add data distributions, and values for descriptors, and/or statistical significance as shown in Fig. 7.

## 6 Automatic Pattern Detection and Verification

SmartExplore has fully automatic exploration support, such as reliability analysis or table ordering, to increase trust in the findings.

### 6.1 Pattern-based Layout

The perception of patterns in the SmarTable depends on the ordering of rows and columns. Therefore, SmartExplore implements automatic sorting strategies to reveal these patterns. Since SmartExplore allows visualizing numerical, categorical, and binary dimensions, our internal heuristics can automatically select the correct distance functions for the involved data type (-combination). For all sorting strategies, similar descriptors, record groups, and dimensions should be placed close to each other. However, finding a good table reordering can be seen as an optimization problem in which row- and column positions can be freely changed without affecting underlying data interpretation [4]. Yet, finding an ‘optimal’ solution is often computationally impossible or reveals the problem that reordering algorithms are inherently designed to foster the visual appearance of *one* visual pattern [4].

#### Automatic Sorting of Groups and Dimensions

In SmartExplore, we can luckily restrict our search for an appropriate reordering algorithm to those approaches that are known to promote the visual patterns presented in Section 4.3. Hence, at least three options are possible: (a) the Barycenter reordering [41], the Bond-Energy algorithm (BEA) [42], or Correspondence Analysis (CA) algorithms [26]. Barycenter and CA are both fast algorithms but are designed to *only* retrieve groupings around the diagonal. CA implements a Singular Value Decomposition, which is not applicable if distances are less discriminative (i.e., binary or categorical dimensions). We decided to implement the BEA algorithm as it is a more conservative approach. This algorithm internally optimizes the ‘measure of effectiveness’, which fosters the visual appearance of groups independent of their relative location to the main diagonal. Moreover, these groups do not necessarily have to have a quadratic shape, but can also be rectangular.

#### Automatic Sorting of Dimensions

SmartExplore implements two strategies to automatically sort dimensions within a subspace based on the given ordering of dimensions. Same as before, this ordering is automatically applied until the user changed the ordering manually.

**Sorting by average descriptor.** The first approach sorts all dimensions of a subspace ascendingly by the average descriptor per dimension. An example can be found in Fig. 1. While this sorting can be applied to both normalizing strategies, it is most useful when the dimensions are normalized across the subspace. As a result, users can quickly see which dimensions generally have higher/lower measures.

**Sorting by visual similarity.** The second strategy sorts the dimension by visual similarity. First, SmartExplore computes a distance matrix by all pairs of dimensions within a subspace. To do so, the previously introduced distance measures based on the (weighted) Euclidean distance are used. Afterward, we compute a one-dimensional multi-dimensional scaling projection of the distance matrix, similar to proposed in [30]. We ignore the actual position in the one-dimensional layout but use the ordering of the projected dimensions. For stacked grouping, users can select which parts to consider for the layout: the base measure, the stacked measures, or a combination of both.

## 6.2 Automatic Pattern Detection

In Section 5.2, we describe how users can select a dimension of interest and highlight all dimensions that are visually similar. This user-guided analysis is particularly interesting for the application of specific patterns of interest. However, in most applications, users are primarily interested in linear correlations, clusters, and outliers as introduced by our pattern topology. SMARTEXPLORE supports users in automatically identifying these patterns: For each pattern-type, we defined a *template* describing the 'optimal' pattern for a single dimension. These templates correspond to the examples of the pattern topology, as shown in Figures 4 and 5. We adapt the size of the pattern to the number of rows in the (stacked) SMARTABLE. For patterns like the outliers in Fig. 5 (e), we iterate the position of the pattern (here: outlier) through all rows of the dimension. Finally, the different templates for each pattern are matched against each dimension in the dataset - analog to the manual similarity search.

## 6.3 Reliability of Visual Patterns

While the visual design supports finding different data patterns, SMARTEXPLORE automatically and transparently supports the analyst in the question *"How reliable are these findings?"* (**R11**).

### Statistical Significance and Visual Representation

Different colors for visualized descriptors naturally indicate that the underlying values are different. However, based on the normalizing strategy and the chosen colormap (e.g., bi-polar), the minimum descriptor (*min*) is mapped to blue and the maximum (*max*) to red. In the visualization, users cannot quantify the difference between *min* and *max* without using the tooltip or table lens. The same is true for all descriptors in-between. Therefore, SMARTEXPLORE *automatically* computes various statistical tests to assess whether differences are statistically significant or not. The following two levels-of-detail are considered:

**S1: Significance of a descriptor.** For every computed descriptor, a statistical test is used to decide whether it is significantly different from the overall dimension. To measure this difference, classical tests are t-tests to compare the mean (descriptor) with the mean of a dimension, Chi$^2$ tests for categorical dimensions, and a binomial test for binary dimensions.

**S2: Significance of a dimension.** To measure the significance of multiple descriptors at the same time, classical tests are an ANOVA for numerical, and a Chi$^2$-test for categorical and binary dimensions. These tests generalize the understanding of **S1** to an entire dimension, but do not indicate the significance of each descriptor.

**Assumption-based selection of statistical test.** Each statistical test relies on different assumptions that need to be fulfilled in order to achieve reliable results. In numerical dimensions, for example, analysts have to check whether the data follows a normal distribution (e.g., using the Kolmogorov-Smirnov test), for variance homogeneity for independent samples (using Levence's test), and sphericity for dependent samples (for ANOVA with rep. measures, using Mauchly test). The same applies to categorical and binary dimensions in which, for example, the sample size has to be taken into account. Following Andy Field [16], there are 11 tests for numerical, three for categorical, and one for binary dimensions that apply to our application. SMARTEXPLORE supports the user by automatically selecting the appropriate test for each dimension. Based on the data type, the (in)dependence of samples, and the significance type **S1** or **S2**, SMARTEXPLORE computes all statistical tests and their assumptions. Appropriate test are selected as proposed by Andy Field [16]. Then the test's $p-value$ is compared to a user-defined $\alpha$ to determine the significance of a dimension or a computed descriptor. The tooltip shows the $p-values$ of all tests and assumptions such that the user can compare their difference and reproduce the system's selection. Users can also manually determine the applied test for a single dimension, a subspace, or globally for the entire dataset.

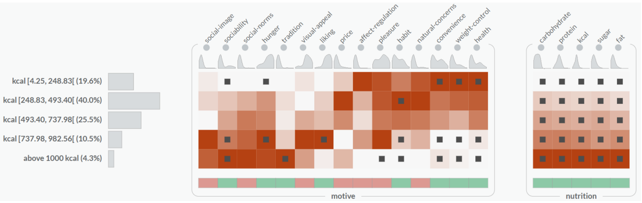


Figure 6: Comparison of nutrition, eating motives, and calories per meal (rows). Meals, rich in calories, are merged into one record group (bottom). Significant dimensions and descriptors (mean) are marked.

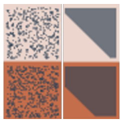
A $p-value$ only informs whether a statistical effect exists; it does not show its magnitude. The appropriate effect size (e.g., Cohen's d, Cramer's V) for the selected tests is also automatically computed.

**Visual representation of statistics.** The statistic results can be added to SMARTABLE. The significance of a descriptor **(S1)** is visualized by an overlay. Users can choose between a dot for significant descriptors (Fig. 6 ▦) and a glyph that uses full size for significant and a smaller size for non-significant descriptors ▦. Applying the first option, users can concentrate on the patterns and use the statistical information as added value. The second option modifies the visual representation such that significant results jump out and users can concentrate on areas with mainly significant descriptors.

To show the significance of a dimension **(S2)**, users can enable a red or green icon below each dimension (Fig. 6 ▦). Also, an adaptive colormap ▬ can be used ▦. Significant dimensions use the full range of colors, non-significant, only the inner part. As a result, users can still perceive differences in the descriptors, but they are visually less dominant as significant ones.

### Missing Values

Missing values are common in many applications and influence the reliability of descriptors. Therefore, the visualization should highlight the areas in the data space which contain missing values and show their proportions. Otherwise, the uncertainty of calculated descriptors is not shown, and the visualization pretends a reliable pattern which does not exist in the underlying data. SMARTEXPLORE supports different visual overlays to show the amount of missing values. For example, the *glyph covering* adds a gray layer on top of the visualized descriptor in order to reduce its expressiveness. The *texture overlay* covers the visualized descriptors with random noise, as used by Buchmüller [7]. Estimating the exact proportion of missing values is not possible. However, it is more intuitive as it seems there are 'holes' in the data, analog to missing values.



## 7 EFFECTIVENESS AND GENERALIZABILITY EVALUATION

We evaluate SMARTEXPLORE for two general criteria: First, its usability and understandability for pattern analysis tasks, and second its generalizability to different datasets and domains.

**Evaluating effectiveness.** To assess the effectiveness and usability of SMARTEXPLORE, we conducted a qualitative expert user study with six participants. Our evaluation process is structured in a multi-stage evaluation process:

(1) We generate a set of 'ground truth findings' from the food dataset derived by two participants, who are familiar with the data due to earlier analysis using established statistics. Both subjects have no far-reaching VA experience, but continuously provided feedback during the development and use SMARTEXPLORE on a regularly basis. We refer to these participants as **E1** and **E2** as they are <u>e</u>xperts in both, the data and SMARTEXPLORE.

(2) We target the usability across *different expertise levels*, by conducting four pair analytics [34] studies with two different user groups. In the first group, two psychologists without VA experience, but good knowledge of the food dataset participated. We refer to these participants as <u>d</u>ata <u>e</u>xperts (**DE1** and **DE2**). In the second group,

Table 1: Overview of expert users and their role during the evaluation.

|  | VA | Dataset(s) | Role | Method |
|---|---|---|---|---|
| **E 1+2**<br>Psych. | novice | `food` | ground truth generation | indep. analysis & feedback |
| **DE 1+2**<br>Psych. | novice | `food` | compare across expertise | pair analytics & interview |
| **VE 1+2**<br>CS | expert | `food`<br>`university` | compare across expertise & data | pair analytics & interview |

two visual analytics experts (PhD students with one to three years of experience), **VE1** and **VE2**, participated. **VE1** and **VE2** did not have knowledge about the `food` dataset. We planned about one hour per pair analytics session per participant and conducted a semi-structured interview for gathering feedback, feature requests, and potential improvements. None of these participants (**DE1+2** and **VE1+2**) has been using SMARTEXPLORE before.

**Evaluating generalizability** In order to showcase SMARTEXPLORE's applicability on datasets of various domains, we also let our VA experts **VE1** and **VE2** analyze the `university- ranking` dataset[1]. This dataset contains the top-1000 universities for the years 2014-2017, ranked according to nine different metrics, such as the quality of education, number of publications and patents. The metrics result in a numerical score, used to derive an overall ranking. As before, we conducted a pair analytics study combined with a semi-structured interview which took 30 minutes in total.

An overview of all user groups, expertise levels, roles, and evaluation methods is shown in Table 1. In the following we will describe the results of each experiment in detail.

### 7.1 Insight Generation

During the last year **E1** and **E2** have been using SMARTEXPLORE in different stages of the implementation. Both experts primarily analyzed the `food` dataset. We are not able to report all findings in this paper, but we will describe interesting usage scenarios and depict the general analysis process of the experts. According to **E1** and **E2**, finding *statistically significant* commonalities among a large set of semantically grouped dimensions, (e.g. eating motives or ingredients), is the most convincing argument for using SMARTEXPLORE.

The experts analyzed how age influences the preference towards certain ingredients (Fig. 1 (A)) ▦. The dimensions are, hereby, normalized within a subspace to find coinciding products that are generally consumed a lot. The experts found (statistically) obvious insights easily, such that *milk*, *small bread*, and *vegetables* are generally consumed more often (dark red colors) than *fish*, *potatoes*, and *pulse* (dark blue colors). Older people (last row) seem to use more milk than younger people, a finding which could be later rejected due to its unreliability ($p-value$ of 0.06). **E1** and **E2** found that there is variance based on the gender, so they created a stacked SMARTABLE (Fig. 1 (B)) ▦. In the group 53-66 years, the amount of vegetables is slightly above average (light red color), but differs strongly for male (less vegetables) and female (more vegetables). The experts made use of our automatic pattern retrieval functionality by selecting this pattern and searched for similar findings. The experts extended the analysis by comparing the age also to different motives (reasons why people consumed a specific meal; Fig. 3). Different normalizing strategies and colormaps were applied. The SMARTABLE illustrates that motives like *convenience*, *hunger*, *affect-regulation*, and *sociability* might be more important for younger people, while older people are more motived by *price*, *tradition*, and *social norms* (top row). The experts also found that, generally, the motives *liking*, *visual-appealing*, and *hunger* are the most common motives. Further analysis results can be found in Fig. 2 ▦ and 6 ▦ in which **E1** and **E2** analyzed the relation between ingredients and nutritions, respectively motives to consume meals with high/low calories.

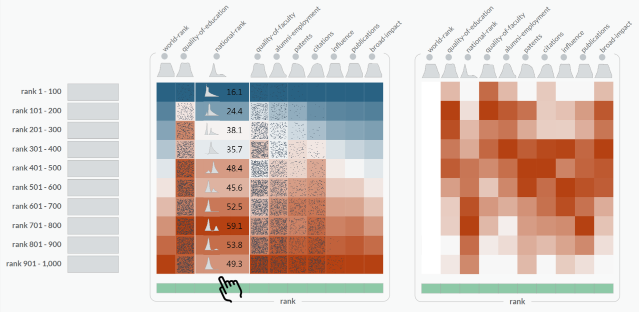


Figure 7: Dataset `university`, grouped by world-rank, and visualized by mean (left) and variance (right subspace). The ten dimensions represent different ranking measures. Left: (blue → good rank and red → low rank); right (white → low and red → high variance). Missing values are shown by noise overlay. Table lens is used to investigate the data distribution and the descriptor of dimension *national rank*.

### 7.2 Comparability across Expertise-levels

To analyze the influence of expertise-levels on the usefulness of SMARTEXPLORE we asked **DE1+2** and **VE1+2** a non-trivial analysis question: *"Which meal type is generally most unhealthy?"* Based on this controversial question, we gave the participants a ten minute introduction and showed them the most important features. All experts showed active interest in our available normalizing strategies, how to interpret particular visual patterns, and asked for the internals of our automatic computation of statistical tests; a circumstance of significant importance, especially in the psychology domain. After understanding that SMARTEXPLORE automatically selects the test based on all assumptions, **DE1** stated that SMARTEXPLORE "[. . . ] not only lets us validate [hypothesis] significantly faster, but also mitigates the problem of choosing accidentally a wrong test".

After all open questions were answered, we asked the participants a second, more open analysis question: *"Which motives and ingredients relate to meals with high, middle, and low calories?"* All participants started analyzing the dataset by grouping the record over the dimension *kcal*. The grouping granularity, however, changed between the different user groups. While **VE1+2** used a grouping with more bins ▦, **DE1+2** created only five bins based on a similar grouping in the literature; Fig. 6 ▦. Independent of the grouping granularity, both participant groups where able to identify a (linear) correlation between *kcal* and all dimensions within the *nutrition* subspace (all statistically significant). **VE1** then merged all record groups with $kcal > 900$ into a single group to remove the distorted distribution of group size. The resulted groups are similar to manual groups of **DE1+2**. Based on this grouping, **VE1** could identify that meals with higher number of calories might be associated with the motives *social-norms*, *hunger*, *tradition*, and *visual-appealing*, while a lower number of calories corresponds to motives like *natural-concerns*, *weight-control*, and *health* (see Fig. 6). Changing the granularity levels of the grouping by *kcal*, the computed descriptors alternated between significant and non-significant. These findings are in line with the 'ground truth' identified by **E1** and **E2**.

Afterward, all participants were motivated to continue analyzing the dataset based on their own interest. **DE1** expanded the search to other dimensions and continued with stacked aggregations separating male vs. female for different meal types. **DE2** started a completely new analysis and looked for patterns w.r.t. stress and mood before and after meals. **VE2** analyzed which ingredients and motives are related to a high body-mass-index. Surprisingly, a small number of participants tried to avoid food and ingredients with a high amount of sugar and calories. As *weight control* is one of the outstanding motives of this record group, **VE2** hypothesized that these participants may be planning or conducting a diet.

[1] Source: http://cwur.org; last accessed: 2018-06-26.

### 7.3 Comparability across Datasets

In a separate session, **VE1+2** started analyzing the `university` dataset. Both VA experts directly applied their experience from the first study and wanted to find out, which aspect correlates mostly with the overall ranking of universities. Therefore, the universities were grouped and binned by their world rank. Fig. 7 shows the mean ▦ and variance ▦ descriptors for all dimensions. Missing values (universities with a *rank* $> 1000$ within one dimension) are visually highlighted with our random noise overlay ▦. Both participants found effortlessly that all of the attributes correlate to the world rank (first dimension). However, there were two observations: (1) the ranking is not linear, and (2) there is a strong variance in all dimension. The dimension *national rank* is visually outstanding as the variance seems to be linearly correlated with the world rank. **VE1** continuously used the tooltip to get the data distribution while **VE2** used the stacked-aggregations to analyze differences in the different years ▦. He found, for example, that the influence of the rank by *patents* changed significantly between 2014 and 2017. Both experts made use of the statistical tests for verification, but relied mainly on the pattern taxonomy, and the distribution overlay ▦ to generate findings.

The reported findings of the `university` dataset are rather an illustrative example than a comprehensive user study. However, we could show that SMARTEXPLORE can be used for other datasets as well and the usefulness is acknowledged by VA experts (see below).

## 8 Discussion and Future Work

In our expert case studies we have shown that SMARTEXPLORE can be applied to various applications. Users with different data and VA expertise are able to identify and understand interesting patterns in HD data. Based on their feedback and our observations during the study, we summarize the following lessons learned:

### Lessons Learned

**Instant applicability through familiar representation.** Both, the **E1+2** and **D1+2** participants have not been using sophisticated VA tools before to find patterns across a large set of dimensions. When we asked them to apply SMARTEXPLORE to their data and give feedback (**E1+2**), and to participate in our study (**D1+2**) the experts showed some skepticism on the usefulness. However, after only a few minutes the familiar representation of the SMARTABLE convinced them instantly to see its usefulness for their own data. Of course, applying SMARTEXPLORE to their own data helped them building a mental relationship between previous findings and the visual patterns. We were able to see that the participants fully understood SMARTEXPLORE by the following observations: (1) During the feedback sessions **E1+2** proposed useful extensions based on the concept of SMARTEXPLORE. For example, they suggested to clone entire subspaces for a comparative analysis using different statistic tests, and initiated the discussion for the automatic reliability analysis. (2) After a short training, **D1+2** directly applied the concepts of SMARTEXPLORE to their own analysis questions. They did not question our design choices but were immediately able to make sense of the visible patterns and explain interesting relationships. Therefore, we conclude that they were able to effectively use SMARTEXPLORE after only a short training phase.

**Findings by automatic support.** We realized that most participants acknowledged the automatic support of SMARTEXPLORE. For example, they liked that similar dimensions are arranged next to each other by default and appropriate statistic test are proposed. As a consequence, the participants were able to spot interesting patterns without any pre-configuration and parameter choices. Once an interesting pattern has been identified, the participants investigated the automatic selections and adjusted the settings.

**Linking to classical approaches.** Even though the layout of the SMARTABLE is quite fixed, sophisticated patterns could be detected by **VE1+2**. Both argued that our design choices along with the automatic support is helpful to identify and explain various patterns. Especially, they liked the possibility to analyze datasets with mixed data types. However, to confirm some of the hypothesis they proposed to transform a subset of the data to other visualization approaches. For example, to see the actual values of records across all dimension (and not just its descriptors), Parallel coordinates are useful.

### Future Work

Although SMARTEXPLORE presents a sophisticated table-based VA system, we identified five areas for future improvements:

**Data types.** We have limited ourselves to datasets with numerical, categorical, and binary dimensions. While the analysis of these mixture datasets is itself challenging, e.g., due to the problematic definition of similarity and aggregations, a broad range of further data types exist. Text-, geo-spatial-, time series-, or relational datasets impose further challenges to both visualization and analytics.

**Layout flexibility.** SMARTEXPLORE's main visualization is a table which borrows the static layout of rows and columns. While it has significant advantages for a broad range of users, we envision a system that lets the user freely change back and forth between known layouts and, e.g., projection-based layouts to facilitate more intuitively high-dimensional similarity assessments.

**Data and analysis provenance.** In SMARTEXPLORE, we present an implicit data provenance approach: All analysis stages are encoded in the URL. However, we found that an explicit gallery or journal view would be highly appreciated by our user group.

**Supporting hypothesis generation.** Within the user study, **VE1** argued that even for unknown datasets users will need an initial hypothesis. **VE1** suggested to show small previous of different aggregations and orderings. Ideally these previous should be sorted and incorporate the user's interaction provenance. **VE2** had a similar idea by proposing to generally highlight relations in the data (e.g., correlation matrix) in order to guide the analysis.

**Trust-building.** One of SMARTEXPLORE's primary contributions is its automatic reliability analysis, which builds trust in the tool and its findings. Further, an algorithmic 'helper', such as subspace clusterings [47] or subspace nearest neighbor search [27], could be explored into (semi-) automated exploration processes.

## 9 Conclusion

Finding and understanding clusters, correlations, and complex patterns in high-dimensional data is a challenging task, especially if the underlying dataset contains a mixture of different data types. With SMARTEXPLORE, we present a fully functional table-based visual analytics technique that combines automatic analysis with user-guided- and purely interactive exploration. In an easy to use interface, our system automatically guides users to interpretable patterns and supports the exploration through semi-automated pattern matching and user invoked reordering. Our interaction concept, based on drag&drop, context-dependent menus, and on-the-fly sliders, allows the user to effectively explore datasets along the record and dimension axis. While some of our approaches are inherent to SMARTEXPLORE's design, we claim that, e.g., our automatic reliability analysis is generalizable to other systems. By means of an expert case studies with users of different expertise, we show that SMARTEXPLORE is effective for a broad audience and application domains.

## Acknowledgments

We wish to thank the German Research Foundation (DFG) for their financial support within the projects A03 & C01 of SFB/Transregio 161. Furthermore, this research was supported by the Federal Ministry of Education and Research within the research project SmartAct (BMBF Grant 01EL1820A). We thank Georges Grinstein, Mennatallah El-Assady, Dirk Streeb, Harald Schupp, Angela Whale, and the anonymous reviewers for their valuable discussion and feedback.

## References

[1] J. Abello and F. van Ham. Matrix zoom: A visual interface to semi-external graphs. In *10th IEEE Symposium on Information Visualization (InfoVis 2004)*, pp. 183–190, 2004. doi: 10.1109/INFVIS.2004.46

[2] G. Albuquerque, M. Eisemann, D. J. Lehmann, H. Theisel, and M. Magnor. Improving the visual analysis of high-dimensional datasets using quality measures. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*, pp. 19–26, 2010.

[3] M. Ankerst, S. Berchtold, and D. A. Keim. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis '98)*, pp. 52–60, 1998. doi: 10.1109/INFVIS.1998.729559

[4] M. Behrisch, B. Bach, N. H. Riche, T. Schreck, and J. Fekete. Matrix reordering methods for table and network visualization. *Computer Graphics Forum*, 35(3):693–716, 2016. doi: 10.1111/cgf.12935

[5] M. Behrisch, M. Blumenschein, N. W. Kim, L. Shao, M. El-Assady, J. Fuchs, D. Seebacher, A. Diehl, U. Brandes, H. Pfister, D. Weiskopf, and D. A. Keim. Quality metrics for information visualization. *Computer Graphics Forum*, 37(3):625–662, 2018. doi: 10.1111/cgf.13446

[6] J. Bertin. La graphique et le traitement graphique de l'information. *Nouvelle bibliothèque scientifique, Flammarion*, 1975.

[7] J. Buchmüller, H. Janetzko, G. L. Andrienko, N. V. Andrienko, G. Fuchs, and D. A. Keim. Visual analytics for exploring local impact of air traffic. *Computer Graphics Forum*, 34(3):181–190, 2015. doi: 10.1111/cgf.12630

[8] R. Bujack, T. L. Turton, F. Samsel, C. Ware, D. H. Rogers, and J. Ahrens. The good, the bad, and the ugly: A theoretical framework for the assessment of continuous colormaps. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):923–933, 2018.

[9] N. Cao, D. Gotz, J. Sun, and H. Qu. Dicon: Interactive visual analysis of multidimensional clusters. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2581–2590, 2011.

[10] C. D. Correa, Y. Chan, and K. Ma. A framework for uncertainty-aware visual analytics. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pp. 51–58, 2009. doi: 10.1109/VAST.2009.5332611

[11] T. Cox and A. Cox. *Multidimensional Scaling, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, 2000.

[12] A. Dasgupta and R. Kosara. Pargnostics: Screen-space metrics for parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1017–1026, 2010. doi: 10.1109/TVCG.2010.184

[13] G. P. Ellis and A. J. Dix. The plot, the clutter, the sampling and its lens: occlusion measures for automatic clutter reduction. In *Proceedings of the working conference on advanced visual interfaces, (AVI)*, pp. 266–269, 2006. doi: 10.1145/1133265.1133318

[14] N. Elmqvist, T. Do, H. Goodell, N. Henry, and J. Fekete. Zame: Interactive large-scale graph visualization. In *Proceedings of IEEE Pacific Visualization Symposium*, pp. 215–222, 2008. doi: 10.1109/PACIFICVIS.2008.4475479

[15] S. J. Fernstad, J. Shaw, and J. Johansson. Quality-based guidance for exploratory dimensionality reduction. *Information Visualization*, 12(1):44–64, 2013. doi: 10.1177/1473871612460526

[16] A. Field. *Discovering statistics using IBM SPSS statistics*. sage, 2013.

[17] K. Furmanova, S. Gratzl, H. Stitz, T. Zichner, M. Jaresova, M. Ennemoser, A. Lex, and M. Streit. Taggle: Scalable Visualization of Tabular Data through Aggregation. *arXiv:1712.05944*, 2017.

[18] K. Furmanova, M. Jaresova, B. Kawan, H. Stitz, M. Ennemoser, S. Gratzl, A. Lex, and M. Streit. Taggle: Scaling Table Visualization through Aggregation. *Poster @ IEEE Conference on Information Visualization (InfoVis '17)*, 2017.

[19] Google. Googlesheets, 2018. `https://www.google.com/sheets/about/`, Last accessed on 2018-03-30.

[20] Google. Microsoft excel, spreadsheet software, 2018. `https://products.office.com/en/excel`, Last accessed on 2018-03-30.

[21] J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, pp. 857–871, 1971.

[22] S. Gratzl, N. Gehlenborg, A. Lex, H. Pfister, and M. Streit. Domino: Extracting, comparing, and manipulating subsets across multiple tabular datasets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2023–2032, 2014.

[23] S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit. Lineup: Visual analysis of multi-attribute rankings. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2277–2286, 2013. doi: 10.1109/TVCG.2013.173

[24] M. Greenacre. *Correspondence analysis in practice*. CRC press, 2017.

[25] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques, 3rd edition*. Morgan Kaufmann, 2011.

[26] M. O. Hill. Correspondence analysis: A neglected multivariate method. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 23(3):pp. 340–354, 1974.

[27] M. Hund, M. Behrisch, I. Färber, M. Sedlmair, T. Schreck, T. Seidl, and D. Keim. Subspace Nearest Neighbor Search - Problem Statement, Approaches, and Discussion. In *Similarity Search and Applications*, pp. 307–313. Springer International Publishing, 2015. doi: 10.1007/978-3-319-25087-8_29

[28] I. Hur and J. S. Yi. Simulsort: Multivariate data exploration through an enhanced sorting technique. In *Proceedings of Human-Computer Interaction. Novel Interaction Methods and Techniques*, pp. 684–693, 2009. doi: 10.1007/978-3-642-02577-8_75

[29] A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(2):69–91, 1985. doi: 10.1007/BF01898350

[30] D. Jäckle, F. Fischer, T. Schreck, and D. A. Keim. Temporal MDS Plots for Analysis of Multivariate Data. *IEEE Transactions on Visualization and Computer Graphics*, 22(01), 2016. doi: 10.1109/TVCG.2015.2467553

[31] W. Jentner, D. Sacha, F. Stoffel, G. Ellis, L. Zhang, and D. A. Keim. Making machine intelligence less scary for criminal analysts: reflections on designing a visual comparative case analysis tool. *The Visual Computer*, pp. 1–17, 2018. doi: 10.1007/s00371-018-1483-0

[32] S. Johansson, M. Jern, and J. Johansson. Interactive quantification of categorical variables in mixed data sets. In *Proceedings of the International Conference on Information Visualisation (IV)*, pp. 3–10, 2008. doi: 10.1109/IV.2008.33

[33] I. T. Jolliffe. *Principal Component Analysis*. Springer, New York, 1986. doi: 10.1007/978-1-4757-1904-8

[34] L. T. Kaastra and B. D. Fisher. Field experiment methodology for pair analytics. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization (BELIV)*, pp. 152–159, 2014. doi: 10.1145/2669557.2669572

[35] E. J. Keogh, L. Wei, X. Xi, S. Lonardi, J. Shieh, and S. Sirowy. Intelligent icons: Integrating lite-weight data mining and visualization into GUI operating systems. In *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM)*, pp. 912–916, 2006. doi: 10.1109/ICDM.2006.90

[36] S. Konecni, J. Zhou, and G. Grinstein. Advanced interactions with heatmaps for analyzing high-dimensional datasets. 2010.

[37] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister. Upset: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1983–1992, 2014. doi: 10.1109/TVCG.2014.2346248

[38] A. Lex, M. Streit, C. Partl, K. Kashofer, and D. Schmalstieg. Comparative analysis of multidimensional, quantitative data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1027–1035, 2010. doi: 10.1109/TVCG.2010.138

[39] A. Lex, M. Streit, H. Schulz, C. Partl, D. Schmalstieg, P. J. Park, and N. Gehlenborg. Stratomex: Visual analysis of large-scale heterogeneous genomics data for cancer subtype characterization. *Computer Graphics Forum*, 31(3):1175–1184, 2012. doi: 10.1111/j.1467-8659.2012.03110.x

[40] S. Liu, D. Maljovec, B. Wang, P. Bremer, and V. Pascucci. Visualizing high-dimensional data: Advances in the past decade. *IEEE Transactions on Visualization and Computer Graphics*, 23(3):1249–1268, 2017. doi: 10.1109/TVCG.2016.2640960

[41] E. Mäkinen and H. Siirtola. The barycenter heuristic and the reorderable matrix. *Informatica (Slovenia)*, 29(3):357–364, 2005.

[42] W. T. McCormick, S. B. Deutsch, J. J. Martin, and P. J. Schweitzer. Identification of data structures and relationships by matrix reordering techniques. Technical report, DTIC Document, 1969.

[43] K. T. McDonnell and K. Mueller. Illustrative parallel coordinates.

*Computer Graphics Forum*, 27(3):1031–1038, 2008. doi: 10.1111/j.1467-8659.2008.01239.x

[44] Microsoft. Power bi software, Interactive Data Visualization BI Tools, 2018. `https://powerbi.microsoft.com/en-us/`, Last accessed on 2018-03-30.

[45] S. Mittelstädt, D. Jäckle, F. Stoffel, and D. A. Keim. Colorcat: Guided design of colormaps for combined analysis tasks. In *Eurographics Conference on Visualization (EuroVis)-Short Papers. The Eurographics Association*, 2015.

[46] C. Niederer, H. Stitz, R. Hourieh, F. Grassinger, W. Aigner, and M. Streit. TACO: Visualizing Changes in Tables Over Time. *IEEE Transactions on Visualization and Computer Graphics*, 2017.

[47] L. Parsons, E. Haque, and H. Liu. Subspace Clustering for High Dimensional Data: A Review. *ACM SIGKDD Explorations Newsletter*, 6(1):90–105, 2004. doi: 10.1145/1007730.1007731

[48] W. Peng, M. O. Ward, and E. A. Rundensteiner. Clutter reduction in multi-dimensional data visualization using dimension reordering. In *Proceedings of the 10th IEEE Symposium on Information Visualization (InfoVis 2004)*, pp. 89–96, 2004. doi: 10.1109/INFVIS.2004.15

[49] C. Perin, P. Dragicevic, and J.-D. Fekete. Revisiting bertin matrices: New interactions for crafting tabular visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2082–2091, 2014.

[50] R. Rao and S. K. Card. The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pp. 318–322, 1994. doi: 10.1145/191666.191776

[51] G. E. Rosario, E. A. Rundensteiner, D. C. Brown, M. O. Ward, and S. Huang. Mapping nominal values to numbers for effective visualization. *Information Visualization*, 3(2):80–95, 2004. doi: 10.1057/palgrave.ivs.9500072

[52] SAS. Jmp software, JMP Software from SAS, 2018. `https://www.jmp.com`, Last accessed on 2018-03-30.

[53] J. H. Siegel, E. J. Farrell, R. M. Goldwyn, and H. P. Friedman. The surgical implications of physiologic patterns in myocardial infarction shock. *Surgery*, 72(1):126–141, 1972.

[54] H. Siirtola and E. Mäkinen. Constructing and reconstructing the reorderable matrix. *Information Visualization*, 4(1):32–48, Mar. 2005. doi: 10.1057/palgrave.ivs.9500086

[55] B. W. Silverman. *Density estimation for statistics and data analysis*, vol. 26. CRC press, 1986.

[56] M. Spenke, C. Beilken, and T. Berlage. Focus: The interactive table for product comparison and selection. In *Proceedings of the 9th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pp. 41–50, 1996. doi: 10.1145/237091.237097

[57] Spotfire. Spotfire software, Data Visualization & Analytics Software, 2018. `https://spotfire.tibco.com/`, Last accessed on 2018-03-30.

[58] Tableau. Tableau software, Tableau is business intelligence software that helps people see and understand their data, 2018. `https://www.tableau.com`, Last accessed on 2018-03-30.

[59] A. Tatu, G. Albuquerque, M. Eisemann, J. Schneidewind, H. Theisel, M. Magnor, and D. Keim. Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *IEEE Symposium on Visual Analytics Science and Technology*, pp. 59–66, 2009. doi: 10.1109/VAST.2009.5332628

[60] W. S. Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419, 1952.

[61] E. R. Tufte. *Beautiful evidence*, vol. 1. Graphics Press Cheshire, CT, 2006.

[62] J. Twellmeyer, M. Hutter, M. Behrisch, J. Kohlhammer, and T. Schreck. The visual exploration of aggregate similarity for multi-dimensional clustering. In *In Proceedings of the 6th International Conference on Information Visualization Theory and Applications (IVAPP)*, pp. 40–50, 2015. doi: 10.5220/0005304100400050

[63] S. van den Elzen, D. Holten, J. Blaas, and J. J. van Wijk. Reducing snapshots to points: A visual analytics approach to dynamic network exploration. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):1–10, 2016. doi: 10.1109/TVCG.2015.2468078

[64] P. van der Corput and J. J. van Wijk. Exploring items and features with $\mathrm{i}^\mathrm{f}$, $\mathrm{f}^\mathrm{i}$-tables. *Computer Graphics Forum*, 35(3):31–40, 2016. doi: 10.1111/cgf.12879

[65] L. van der Maaten and G. Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[66] K. Villinger, D. R. Wahl, G. Sproesser, H. T. Schupp, and B. Renner. A visual analysis of the behavioral signature of eating: The case of breakfast. *The European Health Psychologist*, 19(Supp):689, 2017.

[67] D. R. Wahl, K. Villinger, G. Sproesser, H. T. Schupp, and B. Renner. The behavioral signature of snacking : a visual analysis. *The European Health Psychologist*, 19(5):355–357, 2017.

[68] E. Wall, S. Das, R. Chawla, B. Kalidindi, E. T. Brown, and A. Endert. Podium: Ranking data using mixed-initiative visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):288–297, 2018.

[69] B. Wang and K. Mueller. The subspace voyager: Exploring high-dimensional data along a continuum of salient 3d subspaces. *IEEE Transactions on Visualization and Computer Graphics*, 24(2):1204–1222, 2018. doi: 10.1109/TVCG.2017.2672987

[70] J. Xia, F. Ye, W. Chen, Y. Wang, W. Chen, Y. Ma, and A. K. H. Tung. Ldsscanner: Exploratory analysis of low-dimensional structures in high-dimensional datasets. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):236–245, 2018. doi: 10.1109/TVCG.2017.2744098

[71] X. Yuan, D. Ren, Z. Wang, and C. Guo. Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2625–2633, 2013. doi: 10.1109/TVCG.2013.150

[72] H. Zhou, X. Yuan, H. Qu, W. Cui, and B. Chen. Visual clustering in parallel coordinates. *Computer Graphics Forum*, 27(3):1047–1054, 2008. doi: 10.1111/j.1467-8659.2008.01241.x

[73] L. Zhou and C. D. Hansen. A survey of colormaps in visualization. *IEEE Transactions on Visualization and Computer Graphics*, 22(8):2051–2069, 2016.