DOI: 10.1111/cgf.12641
Eurographics Conference on Visualization (EuroVis) 2015
H. Carr, K.-L. Ma, and G. Santucci
(Guest Editors)

Volume 34 (2015), Number 3

Visualnostics: Visual Guidance Pictograms for Analyzing Projections of High-dimensional Data

Dirk J. Lehmann, Fritz Kemmler, Tatsiana Zhyhalava, Marco Kirschke, and Holger Theisel

Visual Computing Group, University of Magdeburg, Germany

Abstract

The visual analysis of multivariate projections is a challenging task, because complex visual structures occur. This causes fatigue or misinterpretations, which distorts the analysis. In fact, the same projection can lead to different analysis results. We provide visual guidance pictograms to improve objectivity of the visual search. A visual guidance pictogram is an iconic visual density map encoding the visual structure of certain data properties. By using them to guide the analysis, structures in the projection can be better understood and mentally linked to properties in the data. We introduce a systematic scheme for designing such pictograms and provide a set of pictograms for standard visual tasks, such as correlation and distribution analysis, for standard projections like scatterplots, RadVis, and Star Coordinates. We conduct a study that compares the visual analysis of real data with and without the support of guidance pictograms. Our tests show that the training effort for a visual search can be decreased and the analysis bias can be reduced by supporting the user's visual search with guidance pictograms.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Picture/Image Generation—Viewing algorithms

1. Introduction

The visual exploration of high-dimensional data is a challenging task. An analyst is faced with high-dimensional structures, complex patterns, different visualization approaches, tools, and interaction techniques. Many analysts can handle this, others might be overburdened by the degree of complexity. In order to ease the visual search, users received more and more additional technical guidance support in recent years, e.g., for the automatic selection of relevant visualizations [SNLH09, TAE*09, AEL*09], the automatic selection of an appropriate visualization technique, or the automatic adjustment of the level of detail [SBS11, SBS*14].

However, providing guidance support w.r.t. the level of experience, objectivity, or consistency of interpretations has not been considered yet. Inexperienced users often have a hard time when visually analyzing data. Beforehand, a lot of costly practice is required to get into data visualization and visual analytics. Inexperienced users and companies might be put off by this necessary initial effort. Even experienced users might be inconsistent and subjective regarding their visual data insights. Clearly, such issues make it difficult for data visualization to exploit its full potential and to be accepted in new sectors and research areas.

To address these drawbacks and to provide further guidance, we introduce *Visualnostics*, i.e., visual guidance pictograms. A visual guidance pictogram is an iconic visual density map that encodes the visual structure of data properties. Users can adjust and correct their visual analysis results with the help of guidance pictograms in an early state of the visual search. Misunderstandings can be reduced, visual structures and data properties can be better mentally related, and (if needed) an analysis landmark can be provided during the visual search. Note that guidance pictograms are neither an alternative to a user-based visual search, nor can a visual search be replaced by them. Instead, they can help to reduce the number of wrong decisions and to quickly familiarize inexperienced users with the data and projection properties, and this way with visual analytics.

A guidance pictogram can be generated by using synthetic data with known data properties, observing this data under projection, and by defining a continuous density distribution in the visualization space. In total, we present:

- an algorithm for designing guidance pictograms,
- a set of guidance pictograms for correlation and distribution analysis of (multivariate) projections,
- an illustration with real data of how guidance pictograms can be used in practice, and

- a user experiment with 14 study participants and 16 common visual tasks to compare a visual search with and without the use of our guidance pictograms.

2. Related Work

There are already a number of guidance concepts that support the user at different stages within the visual analysis process. A recent concept is quality metrics, which map a visualization to a real number. They were mostly designed to automatically detect trends, correlation, and cluster separation in data projections, like scatterplots, RadVis, and parallel coordinates [SNLH09, TAE*09, AEL*09, AEL*10, AEM11, LAE*12]. Graph-theoretic scagnostics with metrics for scatterplots were presented by Wilkinson et al. [WAG05]. Additionally, Pargnostics [DK10] provide a set of screen-space metrics to measure the quality of parallel coordinates regarding their parametrization, such as axis ordering or screen-space resolution. Quality metric techniques are systematized by Bertini et al. [Ber11]. Such metrics can support inexperienced users detecting relevant visualizations. However, one issue still requires attention: what is a convenient interpretation for the remaining visualization space? We provide guidance pictograms to also address this aspect during the visual data analysis process.

An interactive guidance approach has been presented by Scherer et al [SBS11, SBS*14]: free-hand user sketches are compared to the scatterplot space via a trained regression model and a feature-based similarity function, respectively. This concept targets towards experienced users and/or domain experts, since a priori knowledge is required to give reasonable sketch-based input. DimStiller [IMI*10] offers a collection of tools for dimension analysis and dimension reduction integrated within an interactive coherent framework for quickly combining and (drag-an-drop-like) joining together a number of approaches of interest for data analysis purposes. Similar to our concept, the DimStiller also provides an entry point in the data analysis for non-experts, since mathematical foundations is not required to use it. The guidance pictograms of our approach are also indented especially for non-experts to easily get into the visual analytics world. In fact, our approach is "orthogonal" to DimStiller, and both approaches could be used together in order to mutually enhance the user's visual analysis process.

For the generation of guidance pictograms, the following abilities are required: (i) the ability to point out relevant data structures, and (ii) the ability to represent them as abstracted density scalar fields in the visualization domain. Relevant work regarding (i) is given by Daniels et al. who describe data-related parameter properties of the Radial Visualization projection approach [DGRG12]. The influence of its dimension arrangement has been investigated by Di Caro et al. [DCFMFM10]. Nováková et al. [Nv06] investigated relations between spherical structures and their projection in Radial Visualizations [HGM*97] and later regarding data trends [Nv09]. Projection-based abilities to visually identify higher-order relations in biological application scenarios were described by Zhang et al. [ZPW10].

The generation of density fields for (bivariate) data projections, i.e. (ii), is treated as Splatterplots [MG13]. They propose to use a continuous density representation to reduce visual complexity and avoid over-plotting effects if the number of data points grows. We seize their basic idea of visualization abstraction by using density fields, and we extend them to multivariate projections (and to parallel coordinates) in order to provide abstract and characteristic patterns of data structures, in the form of guidance pictograms.

The basic idea for using such pictograms comes from the observation that there are already some visual guidance rules that are implicitly used by analysts. Fig. 1 offers the visual encodings [The00] for bivariate positive and negative correlation in scatterplots and parallel coordinates. Such visual guidance rules sometimes occur in paper work as byproduct or are embedded in a discussion that focuses on a different visual analysis topic. In this work, we systematize the study of visual guidance pictograms for relevant visual tasks and for (multivariate) projection techniques.

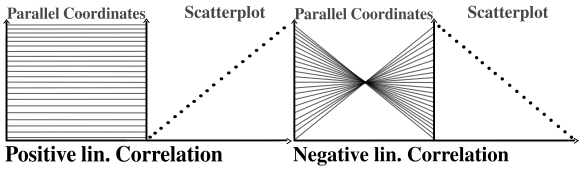

Figure 1: *Some simple and intuitive visual guidance rules.*

3. Focus of our Work

Data properties on different dimension levels are interlocked, built upon, and mutually related. This basic relation is well known and exploited, e.g., by heuristics for subspace clustering [AGGR98]. Moreover, there is a large number of different data visualization approaches which need parametrization, such as axes ordering or anchor point arrangement. Finally, the number of different projections that are required to completely view a dataset grows exponentially with the number of data dimensions. Obviously, it is impossible to cover the number of potential data properties and visual parameter settings within one work. Common visual tasks ask for the visual analysis of trends, correlations, and distributions [TBB*10, SNLH09], to prepare a visual cluster analysis [LAdS12, STMT12] or to support the search for interesting sub-spaces [AGGR98]. Thus, we restrict ourselves to frequently used projection techniques and frequently wanted properties regarding a visual search: To the best of our knowledge, data properties of interest are the analysis of correlations and distributions within bivariate and multivariate projection approaches such as scatterplots, Radial Visualization, and Star Coordinates. In the following, the investigated properties and projection approaches are explained in detail.

4. Data Properties considered for the Design of Visual Guidance Pictograms

We design pictograms as visual templates for *distributions* and *multivariate linear correlations*. Three important data

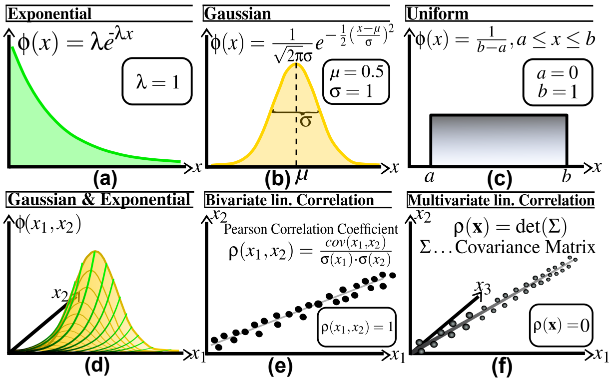

Figure 2: *Data properties considered for the design of visual interpretation rules: distribution properties (a-c), and (multivariate) linear correlations (e-f). In (d), a two-dimensional overlay of distribution is illustrated.*

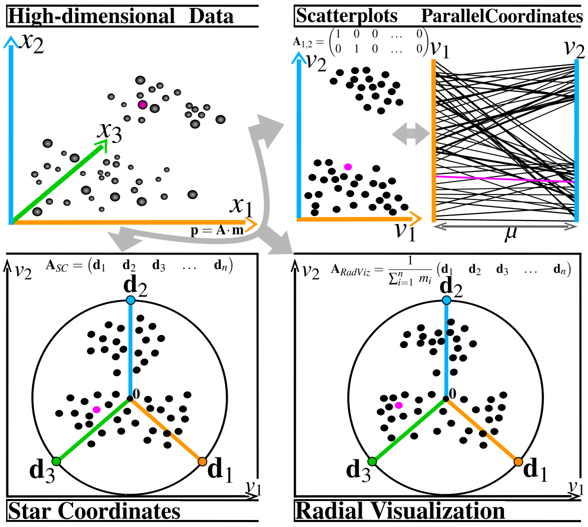

Figure 3: *Our considered projection techniques.*

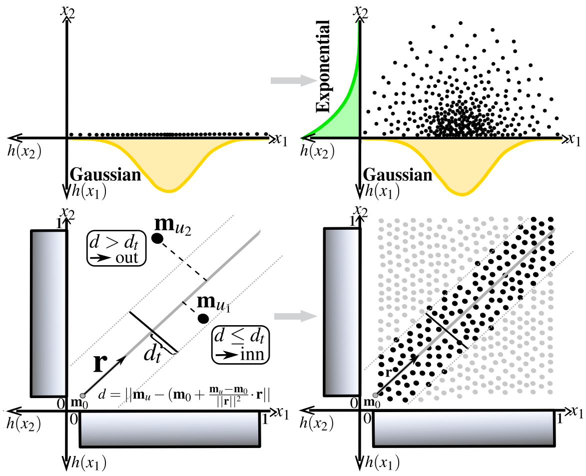

Figure 4: *Schema for modeling the test data*

distributions are considered: the *Exponential*, *Gaussian*, and *Uniform distribution*. They are illustrated in Fig. 2 (a-c). The Exponential typically models time intervals, such as telephone calls or radioactive decay. The Gaussian models the probability behavior of independent events. It is a model for noise or natural processes, and it approximates a special case of the Poisson distribution. The Uniform distribution models events with constant probability, such as the score of a dice, i.e., events which have no intrinsic preference. Thus, concepts from the descriptive statistics, such as "expected value" or "variance", fail for them. Furthermore, we consider multivariate linear correlation to be an important property, since domain experts are usually interested in finding such correlations. As illustrated in Fig. 2 (e-f), correlating dimensions can be (approximately) explained by a line. A bivariate linear correlation between dimension x_1 and x_2 is given by a Pearson Correlation coefficient of $|p(x_1, x_2)| = 1$, a multivariate linear correlation between a number of dimensions $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is given via the covariance matrix $\Sigma(\mathbf{x})$ with $p(\mathbf{x}) = \det(\Sigma) = 0$.

5. Projection Techniques considered for the Design of Visual Guidance Pictograms

We cover the most frequently used projection techniques in the field of visual data analysis. They project an n-dimensional data record $\mathbf{m}$ onto a two-dimensional point $\mathbf{p}$ using a $2 \times n$ projection matrix $\mathbf{A}$, i.e., $\mathbf{p} = \mathbf{A} \cdot \mathbf{m}$. A *scatterplot* [CLN86] is a bivariate orthographic visualization of pairwise data dimensions. An alternative representation of a scatterplot is (bivariate) *parallel coordinates* (PC) [Ins85, Ins09], as illustrated in Fig. 3 (top-right). For this, the axes of the visualization plane are arranged parallel. The component values p_{v_1} and p_{v_2} are connected by a line as visual representation in parallel coordinates of point $\mathbf{p} = (p_{v_1}, p_{v_2})^T$ in the scatterplot (Fig. 3 (top-right)). A *Radial Visualization* (RadVis) [HGM*97] is a multivariate projective projection designed with the aid of two-dimensional anchor points $\mathbf{d}_i$ within a visualization plane. In contrast, *Star Coordinates* [Kan00] use a multivariate affine projection. If the projection is orthographic and multivariate, then *Orthographic Star Coordinates* (OSC) [LT13] are given. Usually, we consider the anchor points $\mathbf{d}_i$ of the multivariate projections to be assigned in a radial layout, as shown by Fig. 3 (bottom), with radius $r = 1$ for the RadVis and the Star Coordinates, and radius $r = \sqrt{2/n}$ for the Orthographic Star Coordinates (cf. [LT13]).

6. Designing Visual Guidance Pictograms

To prepare the design of visual guidance pictograms that exhibit data properties of interest (cf. Sec. 4), generating synthetic high-dimensional datasets (cf. [ALM11]) is necessary. Our approach uses the Box-Muller transform [BM58] in case of a Gaussian, an inverse transform sampling for the Exponential, and the Mersenne Twister [MN98] to generate Uniform distributions, as illustrated in Fig. 4 (top). In contrast, linearly correlating data points are usually scattered along a line $\mathbf{m}_l = \mathbf{m}_0 + t \cdot \mathbf{r}$, with $\mathbf{m}_i$ and $\mathbf{r}$ being n-

dimensional vectors, so that the shortest Euclidean distance $d = d(\mathbf{m}_u, \mathbf{m}_l)$ of record $\mathbf{m}_u$ and line $\mathbf{m}_l$ is small: $d \leq d_t$. Note that d_t equates to the Pearson correlation coefficient, i.e., the smaller d_t, the better modeled is the linear correlation. Hence, our approach generates m uniformly distributed synthetic nD records $\mathbf{m}_{u_i}, i = 1,\ldots,m$ that are kept if $d(\mathbf{m}_u, \mathbf{m}_l) \leq d_t$ applies. Fig. 4 (bottom) illustrates this for the two dimensions x_1 and x_2. Since we are interested in the generation of well-modeled correlations, we use $d_t = 0.2$, $\mathbf{r} = \mathbf{1}$, and $\mathbf{m}_0 = \mathbf{0}$. Our synthetic test data are spatially restricted to a unit hypercube.

Subsequently, our approach projects the synthetic test data $\mathbf{m}_i, i = 1,\ldots,m$ onto points $\mathbf{p}_i$ within the visualization plane. There, for each position $\mathbf{v} = (v_1, v_2)^T$ a continuous approximation of density $\Phi(\mathbf{v})$ is given by

$$\Phi(\mathbf{v}) = \sum_{i=1}^{m} \frac{\Phi(\mathbf{p}_i)}{||\mathbf{v}-\mathbf{p}_i||_2},$$

assuming constant point density $\Phi(\mathbf{p}_i) = 1$. The bias of the density approximation vanishes in the limit $m \to \infty$. In case of parallel coordinates, we exploit the point-line duality to estimate the density. Point-line duality describes that a point $\mathbf{p} = (p_{v_1}, p_\mu)$ in parallel coordinates is associated with a line $\mathbf{L}(\mathbf{p},t) = (0 \quad p_{v_1}/p_\mu)^T + t\cdot(1 \quad (p_\mu - 1)/p_\mu))^T$ in the scatterplot (cf. [LT11]). The density Φ of point $\mathbf{p}$ in parallel coordinates is given by the summed density values along the corresponding line in the scatterplot:

$$\Phi(\mathbf{p}) = \sum_t \Phi(\mathbf{L}(\mathbf{p},t)).$$

Our final guidance pictogram is given by a normalization of density Φ , e.g., by a histogram equalization. For details about normalization, see [LAE*12] (Sec. 4.1.3). The pictogram is a footprint of the data properties

7. Results: Visual Guidance Pictograms

In this section, we provide a set of visual guidance pictograms that are designed as mentioned in Sec. 6. To facilitate an intuitive understanding of the properties that a pictogram represents, we assign a label to each pictogram: $\mathbf{a}_1\mathbf{p}_1\ldots\mathbf{a}_k\mathbf{p}_k$, with $\mathbf{a}_i \in \mathbb{N}^+, \mathbf{p}_i \in$ Property $= \{-\mathbf{c}, \mathbf{c}, \mathbf{u}, \mathbf{e}, \mathbf{n}\}, i = 1,\ldots,k$ where $\mathbf{a}_i$ is the natural number of dimensions, which share the property $\mathbf{p}_i$, and

- "$-\mathbf{c}$" being negative linearly correlated dimensions,
- " **c** " being linearly correlated dimensions,
- " **u** " being uniformly distributed dimensions,
- " **e** " being exponentially distributed dimensions, and
- " **n** " being Gaussian distributed.

The number of considered dimensions per pictogram n_p is given by $n_p = \sum_i \mathbf{a}_i$. For instance, the label "**4c2u**" describes a pictogram of a 6D dataset with 4 multivariate linearly correlated dimensions and 2 further uniformly distributed dimensions. Please note that dimensions being assigned to **c** are based on uniformly distributed synthetic data, and those being assigned to $\mathbf{c}_n$ are based on a Gaussian.

Our resulting guidance pictograms are given in Fig. 5 and 6. Fig. 5 illustrates 48 guidance pictograms for distribution properties (cf. Sec. 4) regarding 2D to 7D (multivariate) projections for the cases "$\mathbf{1\{n,e,u\}1\{n,e,u\}}$" and "$\mathbf{2\ldots4e1\ldots3\{n,u\}}$". For instance, the guidance pictogram for scatterplot of type **2n** shows two Gaussian distributed dimensions that end up in a circular-shaped density flow. More complex appears, for example, the pictogram of a RadVis of type **2e3u**. Here, two exponentially distributed dimensions (green icons) and three uniformly distributed dimensions (blue icons) are shown, which results in a characteristic pentagon-shaped density flow that develops to a rather radial-shaped density peak.

Fig. 6 illustrates 40 guidance pictograms for correlation properties (cf. Sec. 4) regarding 2D to 9D (multivariate) projections for the cases "2$-$**c**", "2$-$**c**1...4**u**" and "2...5**c**1...4**u**". For instance, the guidance pictogram for PC of type **2-c**$_n$ shows two negative correlating dimensions that are Gaussian distributed, which lead to a horizontal "egg timer"-shaped density flow. The pictogram of a RadVis of type **5c2u** represents two noisy dimensions (orange icons) and five multivariate linearly correlating dimensions (green icons) which are all uniformly distributed. The result is a characteristic triangle-shaped density flow. Moreover, the Pearson correlation coefficient can be estimated visually in plots and PCs: the diagonal/vertical d_D of a plot/PC and the diagonal/vertical d_S of visualized values give the ratio $d = d_S/d_D$, from which the Pearson Correlation coefficient $p(d)$ can be visually estimated by the use of the formulas in Fig. 6 (top-right). The related p-d diagram illustrates the error of the estimated coefficients by using these formulas (blue, red, green) and the correct Pearson coefficients (black dots). Note that transposing the dimension's properties for the parallel coordinates pictograms equates to a vertical mirroring, and to swapping the x and y-axis for the scatterplot pictograms. Additionally, to exchange the dimension properties by preserving their property cycle in the RadVis and Star Coordinates pictograms equates to a rotation around the visualization center. Finally, for the standard radial layout, the OSC delivers simmilar pictograms to Star Coordinates.

7.1. Discriminability of the Guidance Pictograms

Since some of the guidance pictograms look quite similar, we conducted a pilot experiment to investigate how different pictograms really are. First, we calculated the pairwise similarity of the pictograms by using a normalized cross-correlation analysis [WL07] (see similarity matrices in the additional material). Then, we randomly selected 4 main pictograms, which should be investigated, and each time 4 similar pictograms. Our pictogram test setup consists of:

for RadVis

- <u>**4e3u**</u>, **3e2u** (0.89), **3e3u** (0.94), **4e1u** (0.86), **4e2u** (0.93),

- <u>**5c3u**</u>, **4c3u** (0.84), **4c4u** (0.80), **5c2u** (0.78), **5c4u** (0.86),

and for Star Coordinates

- <u>**3e2u**</u>, **3e3n** (0.86), **3e3u** (0.90), **4e2n** (0.91), **4e2u** (0.84),

- <u>**4c3u**</u>, **3c3u** (0.90), **3c4u** (0.88), **5c3u** (0.91), **5c4u** (0.85).

A main pictogram (underlined) was presented to a participant for 30 seconds. Then, the participant was distracted for

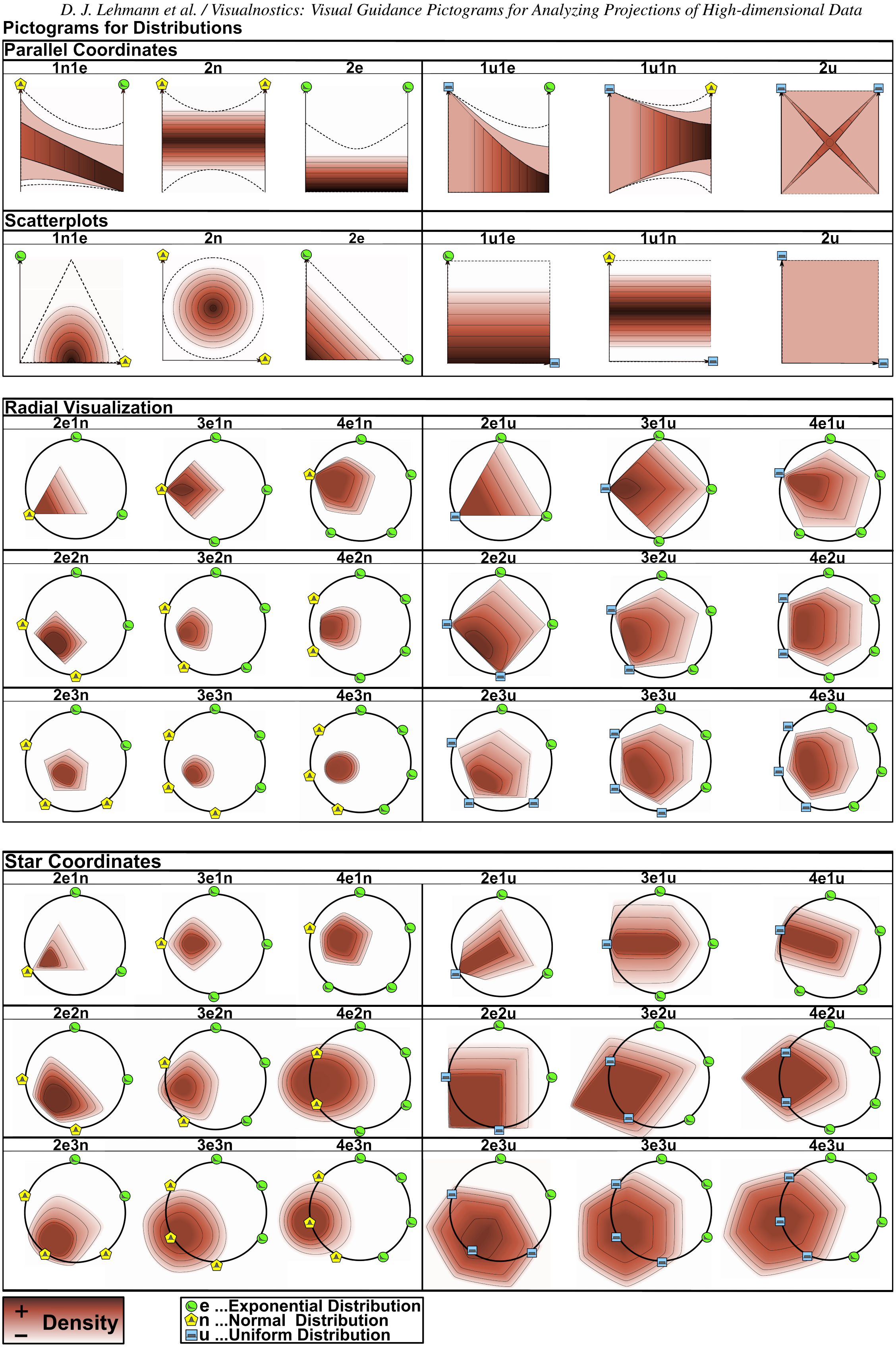

Figure 5: *Visual guidance pictograms for the distribution analysis*

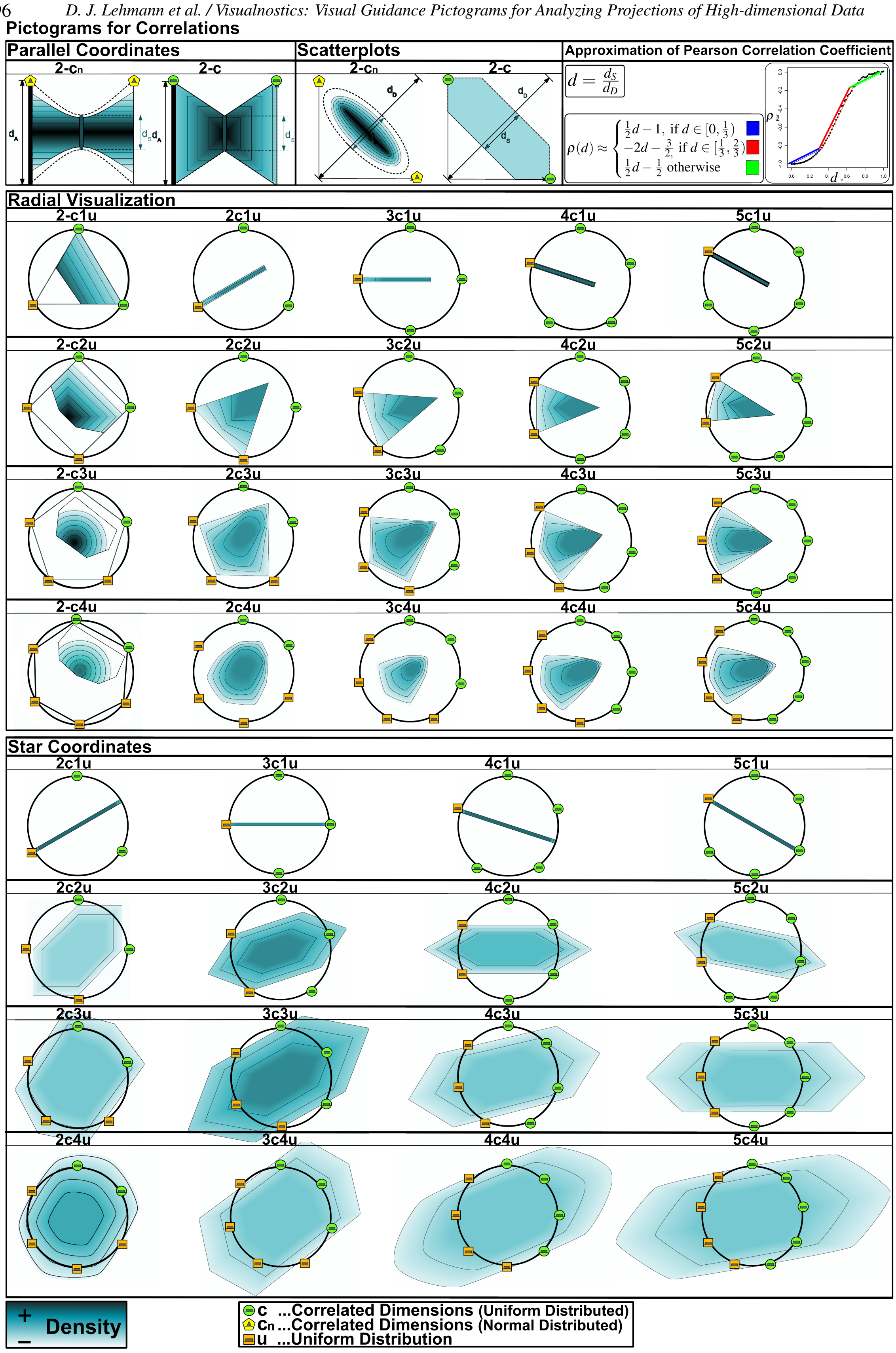

Figure 6: *Visual guidance pictograms for the correlation analysis*

one minute by showing some videos (with audio). Finally, the similar pictograms (similarity represented in brackets) and the main pictogram were presented in random order and a participant had to point out the correct main pictogram.

5 graduated non-expert study participants attended. All the 4 main pictograms were successively tested, i.e., we got 20 iterations. 18 out of 20 were correctly identified, which gives a hit ratio of 90%. Thus, the pictograms seem to be most frequently discriminable for a regular user. Restrictions and limitations are discussed in the following.

7.2. Limitations

Uniqueness: A projection from high-dimensional space to the visualization space is not bijective: a number of different patterns might project onto the same visual structure. This means there is no one-to-one relation from pictograms to data properties and thus false positives might occur. From empirical observation, we found a 9 % rate of false positives in our application test scenario (cf. Sec. 9).
Completeness: The presented pictograms do not completely cover the space of data properties and visualization parameter configurations, since it is not possible to take all aspects into account within a single work. Thus, we focus on projection techniques and data properties that we consider as currently the most important.

8. User Experiment

In order to evaluate if the use of our guidance pictograms improves the visual search in a real life scenario, we conducted a controlled experiment in the laboratory, inspired by [TBB*10,Lea12,LAdS12,STMT12]. We provide a set of visual tasks in which a participant has to find a visualization that fits certain data properties, with and without the support of pictograms. The most obvious design options of a visual task are to provide either a *one-to-one* relation, a *one-to-k* relation, an *m-to-one* relation, or an *m-to-k relation* between a number of m visualizations and a number of k guidance pictograms (see Fig. 7 (top)) with one correct pair that solves the task. Since a participant of our control group will solve the visual tasks without providing the pictograms, a *one-to-one* or a *one-to-k* relation cannot be used because without the pictograms just one visualization remain, which would prevent any vote. In addition, an *m-to-one* relation seems to be not realistic, due to the lack of different (pictogram-related) analysis options, because just one pictogramm is shown. Thus, we decided to use an *m-to-k relation*-based visual task. How a visual task is composed will be explained.

8.1. Composing an Experiment's Visual Task

First, our experiment splits a set of data visualizations into a class $\mathbf{P}$ and a class $\overline{\mathbf{P}}$: A visualization $v_{\overline{\mathbf{P}}}$ of class $\overline{\mathbf{P}}$ is one that does not match to a pictogram. A visualization $v_{\mathbf{P}}$ of class $\mathbf{P}$ is one that matches to a pictogram. Thus, the group $\mathbf{P}$ is directly related to the set of guidance pictograms. How do we check whether a visualization matches to a pictogram? We statistically check the properties of those dimensions that are visualized with the properties that are represented by a pictogram (see Fig. 7 (1)). In terms of distribution properties, we check all combinations (Exponential, Uniform, Gaussian) by using a Chi-square Test [Ken70]. In terms of (multivariate) linear correlations, we apply the Pearson correlation coefficient ρ for the dimensions of bivariate visualizations and for multivariate visualizations we use the determinant of the covariance matrix Σ. Since a perfect correlation can be rarely observed, due to noise in the data, we define a correlation if $|\rho| \geq 0.9$ or $\det(\Sigma) \leq 0.1$.

In order to design a visual task, w.r.t. a visualization technique of a certain dimensionality, our experiment randomly selects a number m of four views from set $\overline{\mathbf{P}}$ (see Fig. 7 (2a), blue bordered) and one view $v_{\mathbf{P}}$ of set $\mathbf{P}$ (see Fig. 7 (2b), orange bordered), yielding five views, presented in random order. Following [Hea96], using more than five visualizations is not advisable. Additionally, our experiment selects the related pictogram of view $v_{\mathbf{P}}$ (see Fig. 7 (3a), orange bordered) and the number of k pictograms with the same dimensionality and the same visualization technique (see Fig. 7 (3b), green bordered). They represent all possible pictogram options in order to make the user's comparison challenging. In this *m-to-k relation*-based visual task setup (see Fig. 7 (4)), only one pair of view and pictogram is related, which is the correct answer of the task; the other views/pictograms are visual noise. The visual task text shown to the user is tailored to this pair, e.g., "Which of these RadVis shows a correlation of 2 dimensions". This setup gives a complex but realistic scenario, because a set of views is presented that also occurs in practice, i.e., when using a scatterplot matrix (SPLOM) or multi-view technique. Additionally, a set of options (i.e. pictograms) is presented, which might explain a view and a certain task (i.e. find fitting pairs) has to be solved. If it is possible to prove that the hit ratio increases by using guidance pictograms even for such a complex visual task scenario, then one could argue that guidance pictograms might also be helpful in simpler scenarios, which is the underlying idea of our visual task setup.

8.2. Characteristics of the Experiment

For our experiment, we used the following datasets: *Yeast* [HN96] (10 dimensions, 1484 records), *WDBC* [FA] (32 dimensions, 569 records), *Cars* [FA] (33 dimensions, 7755 records), *Subway* [FA] (104 dimensions, 423 records), and *Communities* [FA] (128 dimensions, 1994 records). Since millions of different projections exist – due to different projection approaches and visual parameter settings – we randomly generated a set of 12700 projections, which were then assigned to class $\mathbf{P}$ or $\overline{\mathbf{P}}$. Based on this set, we provide 16 visual tasks w.r.t. the investigated visualization techniques. Fig. 8 exemplarily illustrates 2 tasks that we posed (see the additional material for all used visual tasks). The tasks were presented in random order.

We acquired 14 graduated study participants, which were split into two groups of 7 participants. The guidance pictograms of a visual task (=guidance pictogram view) were

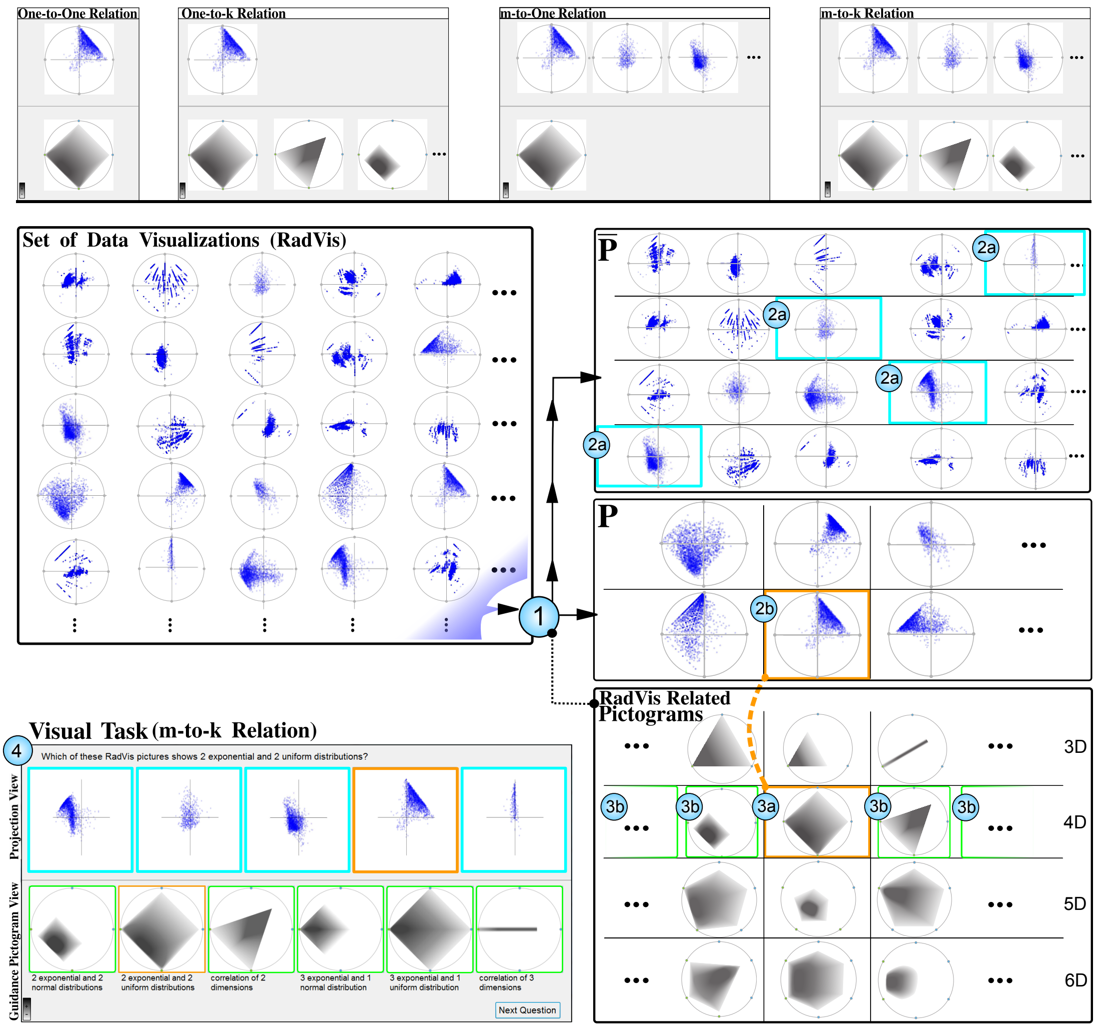

Figure 7: *Composing of a Visual Task: illustrated for RadVis and a number of dimensions $n = 4$.*

only presented to group A, but not to group B. We asked each participant about *visualization experience* on a scale of 1 (low) to 5 (high), *gender*, and *age*. Both groups have similar mean age (A: 29.5 years, B: 27.7 years) and similar experience (A: 3.42, B: 3.14).

8.3. Results of the Experiment

Fig. 9 presents the results of our experiment: the *hit ratio* is the ratio of the number of correct answers divided by the total number of tasks, i.e. 16. The *mean hit ratio* is the mean value of the hit ratio of a group. It can be seen that the mean hit ratio of group A with 0.73 ($\sigma_A = 0.12$) is larger than the mean hit ratio of the control group B with 0.54 ($\sigma_B = 0.19$). In order to check the significance of this difference, we applied a statistical Welch test [JS14] to our results. The outcome is a p-value of $p = 0.04$, meaning that the difference between both groups is significant, regarding a level of significance of 95%. Furthermore, the *mean time* is the average time that a participant needs to conduct a visual task. The mean value of the mean time for group A with 46.9 s ($\sigma_A = 17.1$ s) is a bit larger than for group B with 41.7 s ($\sigma_B = 14.2$ s). We conducted another Welch test. It turns out that the difference regarding the mean time between both groups is not significant with $p = 0.54$, again regarding a significance level of 95%. It remains unclear whether a slowdown is a characteristic property when guidance pictograms are used, but it is still an option, which can be seen in Fig. 9 (middle). Since the participants continuously compare the pictograms with the data projections, this conclusion seems to be plausible. The bottom line is that our guidance pictograms significantly improve the hit ratio for a visual search in our experiment. The mean improvement of the hit ratio by the support of guidance pictograms is +19%.

9. Empirical Validation of Visual Guidance Pictograms

We thoroughly visually searched the above used set of projections for patterns being similar to our pictograms. Especially for multivariate projections, it is known – as an ef-

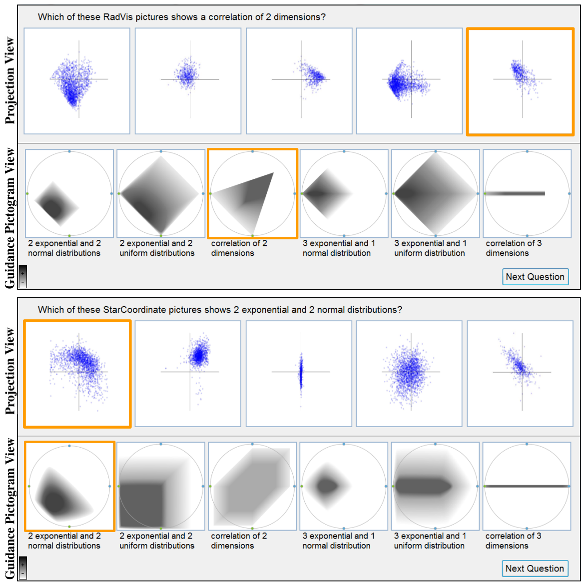

Figure 8: *Examples of visual tasks of the user experiment.*

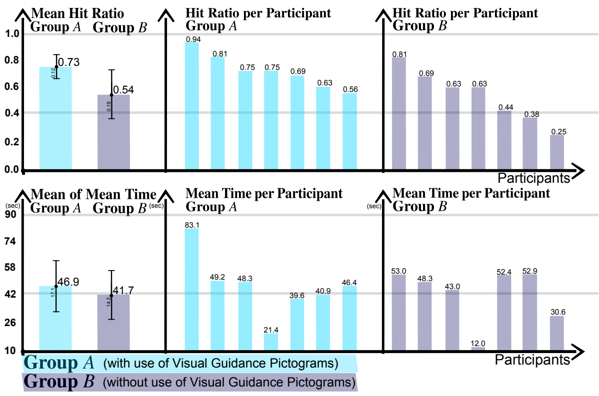

Figure 9: *Results of the user experiment.*

fect of the curse of dimensionality – that the visual contrast decreases with a growing number of dimensions [BGRS99, HAK00]. The effect is that the visualizations often show centered "blob"-like shapes, which convey no information. Such projections were rejected. From the remaining projections, we selected those that qualitatively match with any of our guidance pictograms with same dimensionality. For them, we checked if the data properties match the related visual guidance pictogram properties by using the same approach we used to assign the views to class **P** and class $\overline{\mathbf{P}}$ (cf. Sec. 8). A projection was correctly selected if its properties match with the properties that a pictogram represents. Fig. 10 illustrates samples of real data projections compared to related guidance pictograms that we found during the visual search (some more can be found in the additional material). It turned out that a selected projection does not contain the expected pictogram-related properties in about 9% of the cases, i.e., a 9% false positive rating, but more than 90% of the projections were correctly selected.

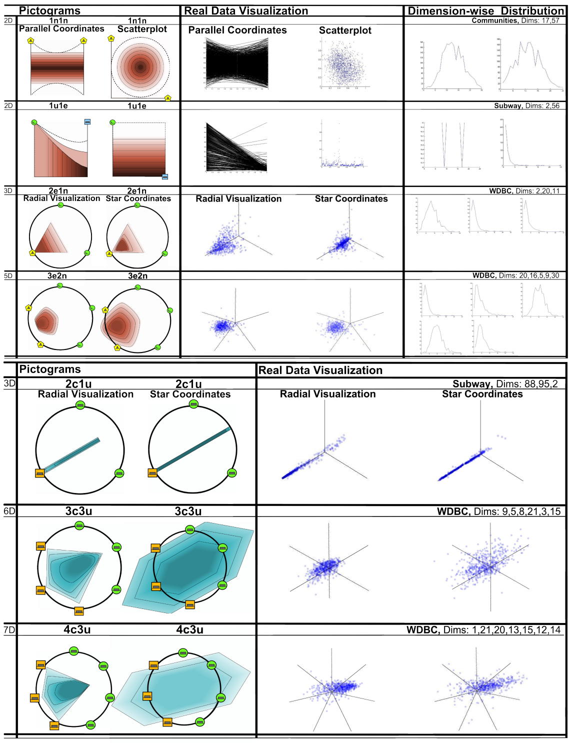

Figure 10: *Visual guidance pictograms compared with selected data projections for distribution properties (top) and correlation properties (bottom).*

10. Conclusion and Future Work

We introduced Visualnostics, the first approach for visual guidance pictograms. Our investigation shows that they are a promising concept to give inexperienced users a starting point or to facilitate a more robust visual search regarding the level of comparability and objectivity between analysis results of different users. For the future we plan to design pictograms that are related to dimension reduction techniques, such as PCA or MDS, and to consider further visualization schemes, e.g., pixel-oriented techniques, and further data properties.

Acknowledgements

The work was supported by DFG grant TH 692/6-1.

References

[AEL*09] Albuquerque G., Eisemann M., Lehmann D. J., Theisel H., Magnor M.: Quality-based visualization matrices. In *VMV* (2009), pp. 341–350. 1, 2

[AEL*10] Albuquerque G., Eisemann M., Lehmann D. J., Theisel H., Magnor M.: Improving the visual analysis of high-dimensional datasets using quality measures. In *IEEE VAST* (2010), pp. 19–26. 2

[AEM11] Albuquerque G., Eisemann M., Magnor M. A.: Perception-based visual quality measures. *IEEE VAST* (2011), 13 – 20. 2

[AGGR98] Agrawal R., Gehrke J., Gunopulos D., RAGHAVAN P.: Automatic subspace clustering of high dimensional data for data mining applications. *SIGMOD Rec. 27*, 2 (1998), 94 – 105. 2

[ALM11] ALBUQUERQUE G., LÖWE T., MAGNOR M.: Synthetic generation of high-dimensional datasets. *IEEE InfoVis* (2011). 3

[Ber11] BERTINI E.: Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE TVCG 17* (2011), 2203–2212. 2

[BGRS99] BEYER K. S., GOLDSTEIN J., RAMAKRISHNAN R., SHAFT U.: When is "nearest neighbor" meaningful? In *Proc. of the 7th International Conference on Database Theory* (1999), pp. 217–235. 9

[BM58] BOX G. E. P., MULLER M. E.: A note on the generation of random normal deviates. *The Annals of Mathematical Statistics 29*, 2 (1958), 610 – 611. 3

[CLN86] CARR D. B., LITTLEFIELD R. J., NICHLOSON W. L.: Scatterplot matrix techniques for large n. *Proc. of 17th Symposium on the Interface of Computer Sciences and Statistics on Computer Science and Statistics* (1986), 297 – 306. 3

[DCFMFM10] DI CARO L., FRIAS-MARTINEZ V., FRIAS-MARTINEZ E.: Analyzing the role of dimension arrangement for data visualization in radviz. In *Proc. of 14th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining - Volume Part II* (2010), pp. 125–132. 2

[DGRG12] DANIELS K. M., GRINSTEIN G., RUSSELL A., GLIDDEN M.: Properties of normalized radial visualizations. *IEEE InfoVis*, 4 (2012), 273 – 300. 2

[DK10] DASGUPTA A., KOSARA R.: Pargnostics: Screen-space metrics for parallel coordinates. *IEEE TVCG 16*, 6 (2010), 1017–1026. 2

[FA] FRANK A., ASUNCION A.: Uci machine learning repository, university of california, irvine, school of information and computer sciences, 2010. 7

[HAK00] HINNEBURG A., AGGARWAL C. C., KEIM D. A.: What is the nearest neighbor in high dimensional spaces? In *VLDB '00: Proceedings of the 26th International Conference on Very Large Data Bases* (2000), Morgan Kaufmann Publishers Inc., pp. 506–515. 9

[Hea96] HEALEY C. G.: Choosing effective colours for data visualization. In *IEEE Visualization* (1996), p. 263ff. 7

[HGM*97] HOFFMAN P., GRINSTEIN G., MARX K., GROSSE I., STANLEY E.: Dna visual and analytic data mining. In *Proc. of the 8th conference on Visualization* (1997), pp. 437–ff. 2, 3

[HN96] HORTON P., NAKAI K.: A probabilistic classification system for predicting the cellular localization sites of proteins. In *Proc. of the Fourth International Conference on Intelligent Systems for Molecular Biology* (1996), pp. 109 – 115. 7

[IMI*10] INGRAM S., MUNZNER T., IRVINE V., TORY M., BERGNER S., MÖLLER T.: Dimstiller: Workflows for dimensional analysis and reduction. In *IEEE VAST* (2010). 2

[Ins85] INSELBERG A.: The plane with parallel coordinates. *The Visual Computer 1*, 2 (1985), 69–91. 3

[Ins09] INSELBERG A.: Parallel coordinates. *Springer Berlin* (2009). 3

[JS14] JAN S.-L., SHIEH G.: Sample size determinations for welch's test in one-way heteroscedastic anova. *British Journal of Mathematical & Statistical Psychology 67* (2014), 72 – 93. 8

[Kan00] KANDOGAN. E.: Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions. *IEEE InfoVis* (2000). 3

[Ken70] KENDALL M. G.: *Rank Correlation Methods*. Griffin, London, England, 1970. 7

[LAdS12] LEWIS J., ACKERMAN M., DE SA V.: Human cluster evaluation and formal quality measures: A comparative study. *Proc. 34th Conf. of Cognitive Science Society* (2012). 2, 7

[LAE*12] LEHMANN D. J., ALBUQUERQUE G., EISEMANN M., MAGNOR M., THEISEL H.: Selecting coherent and relevant plots in large scatterplot matrices. *CGF* (2012). 2, 4

[Lea12] LAM H., ET AL.: Empirical studies in information visualization: Seven scenarios. *IEEE TVCG 18*, 9 (2012). 7

[LT11] LEHMANN D. J., THEISEL H.: Features in Continuous Parallel Coordinates. *IEEE TVCG 17*, 12 (2011), 1912–1921. 4

[LT13] LEHMANN D. J., THEISEL H.: Orthographic star coordinates. *IEEE TVCG 19*, 12 (2013), 2615–2624. 3

[MG13] MAYORGA A., GLEICHER M.: Splatterplots: Overcoming overdraw in scatter plots. *IEEE TVCG 19*, 9 (2013). 2

[MN98] MATSUMOTO M., NISHIMURA T.: Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. Model. Comput. Simul. 8*, 1 (Jan. 1998), 3–30. 3

[Nv06] NOVÁKOVÁ L., ŠTEPÁNKOVÁ O.: Multidimensional clusters in radviz. In *Proc. of the 6th WSEAS International Conference on Simulation, Modelling and Optimization* (2006), pp. 470–475. 2

[Nv09] NOVÁKOVÁ L., ŠTĚPÁNKOVÁ O.: Visualization of trends using radviz. In *Proc. of the 18th Symposium on Foundations of Intelligent Systems* (2009), pp. 56–65. 2

[SBS11] SCHERER M., BERNARD J., SCHRECK T.: Retrieval and exploratory search in multivariate research data repositories using regressional features. In *Proc. of 11th Annual International ACM/IEEE joint Conference on Digital Libraries* (2011). 1, 2

[SBS*14] SHAO L., BEHRISCH M., SCHRECK T., VON LANDESBERGER T., SCHERER M., BREMM S., KEIM D. A.: Guided Sketching for Visual Search and Exploration in Large Scatter Plot Spaces. *Proc. EuroVA* (2014). 1, 2

[SNLH09] SIPS M., NEUBERT B., LEWIS J. P., HANRAHAN P.: Selecting good views of high-dimensional data using class consistency. *Proc. EuroVis 28*, 3 (2009), 831 – 838. 1, 2

[STMT12] SEDLMAIR M., TATU A., MUNZNER T., TORY M.: A taxonomy of visual cluster separation factors. *CGF 31* (2012), 1335–1344. 2, 7

[TAE*09] TATU A., ALBUQUERQUE G., EISEMANN M., SCHNEIDEWIND J., THEISEL H., MAGNOR M., KEIM D.: Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *IEEE VAST* (2009), pp. 59–66. 1, 2

[TBB*10] TATU A., BAK P., BERTINI E., KEIM D., SCHNEIDEWIND J.: Visual quality metrics and human perception: an initial study on 2d projections of large multidimensional data. In *Proc. of AVI* (2010). 2, 7

[The00] THEISEL H.: Higher order parallel coordinates. In *VMV* (2000), pp. 415–420. 2

[WAG05] WILKINSON L., ANAND A., GROSSMAN R.: Graph-theoretic scagnostics. *IEEE InfoVis* (2005), 157–164. 2

[WL07] WEI S., LAI S.: Efficient normalized cross correlation based on adaptive multilevel successive elimination. In *8th Asian Conference on Computer Vision* (2007). 4

[ZPW10] ZHANG X., PAN F., WANG W.: Finding high-order correlations in high-dimensional biological data. In *Link Mining: Models, Algorithms and Applications* (2010), pp. 505–534. 2