13 Discussion

This work examined and analyzed patterns in high-dimensional data characterized by discontinuity. Such distance- or density-based patterns are either compact or connected structures. If the structures are compact, inter- versus intracluster distances are relevant. If they are connected, then density relations and neighborhoods play an important role. Here, it was demonstrated that the neighborhood of a point can always be defined based on graph theory. If the neighborhoods are defined based only on distance, then the structure is compact and a Euclidean graph can be used. If the structure is connected, then two subtypes can be deduced from graph theory: direction-based and unidirectional neighborhoods.

In the context of cluster analysis, structures induced by discontinuities lead to natural clusters, as elaborated in chapter 3. The definition of discontinuity in high-dimensional data, presented in chapter 2, enables the generalization of spatial separation, which was described by [Handl et al.] as a third category of clustering criteria [Handl et al., 2005, p. 3202]. Here, in contrast to [Handl et al., 2005], it is argued that there is no distinction between connected and spatially separated structures or between compact and spatially separated structures[76]. Instead, the third category (spatial separation) can be generalized as the prerequisite for natural clusters defined by either compact or connected structures. It was discussed in chapter 3 that, through the application of basic principles founded on graph theory, clustering algorithms usually search for clusters with a predefined structure. However, it is not always clear which structures are sought because the objective functions that are optimized can be mathematically very difficult to understand. An extensive evaluation of the objective functions found in the literature supports this argument and implies two subtypes of structures sought by common clustering algorithms, called direction-based and unidirectional structures. The assumptions put forward in chapter 3 (Figure 3.5) were verified in chapter 10 (Table 10.1) using data sets from the Fundamental Clustering Problems Suite (FCPS). A question arises regarding how one can choose a clustering algorithm that assumes the correct type of cluster structure for a high-dimensional data set without prior knowledge. Here, it is suggested that dimensionality reduction methods for generating (two-dimensional) projections may help solve this problem.

This work has demonstrated that the objective functions used in clustering and projection methods and the quality measures (QMs) used to evaluate them are based on the fundamental distinction between connected and compact structures. The conclusion is that when the task is to achieve a structure-preserving visualization or clustering, the optimization of an objective function could yield misleading results if the underlying structures of the high-dimensional data of interest are unknown. Hence, a completely different approach is required, which, in chapter 7, motivates an extensive review of the application of artificial intelligence in data science. In chapter 7, two interesting concepts are addressed, called self-organization and swarm intelligence. Through self-organization, the irreducible structures of high-dimensional data can emerge, in a process defined as emergence in chapter 7. If properly applied using a swarm of intelligent agents, the approach presented in this work can outperform the optimization of an objective function for the tasks of clustering and dimensionality reduction.

[76] In [Handl et al.], the three categories of clustering criteria were called connectedness, compactness and spatial separation [Handl et al., 2005, p. 3202].

M. C. Thrun, *Projection-Based Clustering through Self-Organization and Swarm Intelligence*, https://doi.org/10.1007/978-3-658-20540-9_13

The Databionic Swarm (DBS) method

> *"[A clustering approach] must be adaptive or exhibit 'plasticity,' possibly allowing for the creation of new clusters, if the data warrants it. On the other hand, if the cluster structures are unstable [...], then it is difficult to ascribe much significance to any particular clustering. This general problem has been called 'the stability/plasticity dilemma' " [Duda et al., 2001, p. 559].*

The work presented herein introduces a clustering algorithm based on a swarm-based projection method combined with a human-understandable visualization technique. In terms of stability and plasticity (chapter 10, Figure 9.1), the Databionic swarm (DBS) framework outperforms common algorithms in clustering tasks on the FCPS.

> *"One source of this dilemma is that with clustering based on a global criterion, every sample can have an influence on the location of a cluster center, regardless of how remote it might be" [Duda et al., 2001, p. 559].*

In contrast to standard approaches, swarm techniques are known for their properties of flexibility and robustness [Bonabeau/Meyer, 2001; Şahin, 2004]. As a swarm technique, DBS clustering is robust with respect to outliers (see chapter 10).
DBS is a flexible and robust clustering framework that consists of three independent modules. The first module is the parameter-free projection method Pswarm, which exploits the concepts of self-organization and emergence, game theory, swarm intelligence and symmetry considerations. The second module is a parameter-free high-dimensional data visualization technique, which generates projected points on a topographic map with hypsometric colors, called the generalized U-matrix. The third module is a clustering method with no sensitive parameters. The clustering can be verified by the visualization and vice versa. The term DBS refers to the method as a whole. DBS enables even a non-professional in the field of data mining to apply its algorithms for visualization and/or clustering to data sets with completely different structures drawn from diverse research fields, simply by downloading the corresponding R package [Thrun, 2017].
Each module of DBS was compared with various competing algorithms, and in the majority of cases, the modules outperformed those algorithms. However, the author of this work concurs with [Coretto/Hennig, 2016] that despite one's best intentions and efforts to conduct fair comparisons of various methods of visualization, projection and clustering, "ultimately it would be good to have comparisons of methods run by researchers who did not have their hand in the design of any of the methods"; this is because "(simulation) studies can always be designed that make any method 'win.' " The author also agrees with [Coretto/Hennig, 2016] that "readers need to make up their own mind about to what extent our study covered situations that are important to them."
With these considerations in mind, DBS was particularly designed to be flexible and to allow the modules to be interchangeable. An expert in the field of data mining may prefer a method with a clear optimization strategy or may not require the entire DBS framework for his/her application. The interchangeability of the modules is useful in such a case. For example, it is possible to use the visualization and clustering module with NeRV instead of Pswarm. Alternatively, a user could cluster a data set using his/her preferred clustering algorithm and then verify the clusters visually using Pswarm and the generalized U-matrix. As another example, a user could use Pswarm and its clustering algorithm with no visualization, by setting the number of clusters with the aid of the dendrogram of the swarm-defined distances. In summary, the

projection based clustering framework proposed here is a user-friendly platform for the visualization of high-dimensional structures and/or for clustering with no sensitive parameters.[77]

Clustering with DBS

> *"[T]he majority of clustering algorithms […] impose a clustering structure on the data set X, even though X may not possess such a structure" [Theodoridis/Koutroumbas, 2009, p. 863].*

Additionally, they may return meaningless results in the absence of natural clusters [Cormack, 1971, pp. 345-346; Handl et al., 2005, p. 3203; Jain/Dubes, 1988, p. 75]. The results presented in this work illustrate that the DBS algorithm does not suffer from these two disadvantages.

The DBS algorithm makes it possible to apply the abstract U-matrix (AU-matrix) [Lötsch/Ultsch, 2014] to a Pswarm projection instead of an emergent self-organizing map (ESOM) projection. The new clustering approach of DBS is defined by using the shortest-path distances [Dijkstra, 1959] of the AU-matrix and a hierarchical approach to clustering. In contrast to swarm-organized projection (SOP) and ESOM, this approach does not require any parameters except the number of clusters and a two-option parameter that specifies the cluster structure as being either compact or connected (see chapter 3 for details). "One of the most difficult decisions to make is the number of clusters" [Everitt et al., 2001, p. 179]. In DBS, the number of clusters and the cluster structure can be easily estimated from a careful examination of the topographic map (by counting the valleys) and with the help of a dendrogram. If the number of clusters and the cluster structure are chosen properly, then the clusters in the topographic map will be well separated by mountains.

It is argued here that DBS clustering should be semi-interactive and requires user supervision to achieve the best possible results. Nevertheless, the results of automatic DBS clustering with no user intervention were also compared with the results of the common clustering algorithms k-means [MacQueen, 1967], partitioning around medoids (PAM) [L. Kaufman/Rousseeuw, 1990], single linkage (SL) [Florek et al., 1951] and spectral clustering [Ng et al., 2002] as well as two state-of-the-art clustering algorithms: the mixture of Gaussians (MoG) method [Fraley/Raftery, 2002] and the Ward algorithm [Ward Jr, 1963]. "Several of the comparative studies […] conclude that Ward's method […] outperforms other hierarchical clustering methods" [Jain/Dubes, 1988, p. 81]. MoG clustering, which is also known as model-based clustering, serves as the reference technique [Bouveyron/Brunet-Saumard, 2014]. Clustering algorithms such as DBscan [Ester et al., 1996] or the ESOM/U-matrix approach [Ultsch et al., 2016a] require additional sensitive and continuous parameters and were omitted from the comparison for that reason. Every clustering algorithm was applied using the default parameter settings and the correct number of clusters. Calculations were performed for 100 trials on the FCPS data sets [Ultsch, 2005c].

The main result achieved in the work presented herein concerns the error rates of the clustering algorithms tested in these trials. As already stated throughout this work, clustering algorithms often predefine the structure of the clusters they seek; e.g., for PAM and k-means, the shape is round, and thus, the structure is compact. Therefore, these algorithms failed on the Chainlink and Atom data sets. In addition, the k-means and spectral clustering algorithms showed large

[77] After this work it was also made available in [Thrun et al., 2017, Thrun/Ultsch, 2017a].

variances in their results on the Hepta and Target data sets. It is known that the k-means algorithm sometimes strongly depends on the order of objects in a data set [L. R. Kaufman/Rousseeuw, 2005, p. 114], which may be the cause of the large variance in the results. This variance was shown through several examples for the spectral clustering algorithm, in which case the results were strongly trial-dependent, even when the parameter settings remain unchanged. The MoG method yielded results of comparably good quality to those of DBS, but it still failed in the case of the Lsun3D data set (in the sense that it showed a large variance) and in the case of the Target data set and its outliers. The MoG approach uses the expectation maximization (EM) algorithm, which is known to be subject to such problems on univariate data sets [Ultsch et al., 2015]. Notably, only "if the underlying distribution comes from a mixture of component densities described by a set of unknown parameters" can it be estimated using MoG approaches [Duda et al., 2001, e.g. p. 581]. This is the case for the FCPS data sets, resulting in high performance of the MoG algorithm. However, natural data sets do not necessarily satisfyhave to meet this assumption. Additionally, the MoG method fails if the dimensionality of the data set is too high (chapter 3).
The automatic DBS clustering showed a small variance in its results and yielded good accuracy for all data sets. In contrast to all other approaches, in every trial in which the clustering accuracy of DBS was worse than that of some other algorithm, its performance could be improved by using the semi-interactive approach. The reason for this ability to improve the results of DBS lies in the main advantage of DBS clustering, namely, the possibility of verifying the clustering results through visualization, as described below. For a clustering algorithm, it is relevant to test for the absence of a cluster structure [Everitt et al., 2001, p. 180], or the clustering tendency [Theodoridis/Koutroumbas, 2009, p. 896]. Usually, tests for the clustering tendency rely on statistical tests [Theodoridis/Koutroumbas, 2009, p. 896]. Unlike other hierarchical clustering algorithms (except for ESOM/U-matrix clustering [Ultsch et al., 2016a]), the DBS algorithm finds no clusters if no natural clusters exist. The clustering tendency is visualized by the generalized U-matrix.

Generalized U-matrix visualization and structure preservation

The technique of producing visualizations in the form of a two-dimensional scatter plot of projected points currently remains the state of the art in cluster analysis (e.g., [Hennig et al., 2015, pp. 119-120, 683-684; Ritter, 2014, p. 223]). However, such a two-dimensional visualization can lead to a misleading interpretation of the underlying structures because the low-dimensional similarities do not completely represent the high-dimensional distances in two dimensions. Two types of error have been identified in the literature (see chapter 5): forward projection error (FPE) and backward projection error (BPE) [Aupetit, 2007; Ultsch/Herrmann, 2005; Venna et al., 2010]. In addition to these errors, this work introduces the concept of structure preservation, which is the preservation of high-dimensional discontinuities such that no points are allowed to intrude into the discontinuity regions of the two dimensional projection.
The FPEs and BPEs were visualized for various projection methods using a two-dimensional gray-scale U-matrix visualization in [Ultsch/Mörchen, 2006]. Such a gray-scale U-matrix is the most commonly used method for displaying dissimilarities in SOMs [K. Tasdemir/Merenyi, 2009, p. 550; Kadim Tasdemir/Merényi, 2012, p. 3]. Here, the idea was to "apply Self-Organizing Map training without changing the best matching unit [prototype] assignment"

[Ultsch/Mörchen, 2006, pp. 3-4] through the transformation of projected points into best matching units, as introduced in this work. Unlike the approach of Ultsch and Mörchen, the newly proposed simplified ESOM (sESOM) algorithm does not require a learning rate, and the cooling scheme is defined by a special neighborhood function based on symmetry considerations, which results in a parameter-free algorithm (cf. [Ultsch/Mörchen, 2006, p. 4]). This makes it possible to visualize SOMs as topographic maps with hypsometric tints [Thrun et al., 2016a], which serves as a basis for a visualization technique that can be applied in combination with any projection method. The third dimension is used to visualize the local BPE and FPE around each projected point in precisely defined height-dependent colors, thereby giving rise to the generalized U-matrix, which is a generalization of the U-map concept [Ultsch, 2003a].

Here, it is argued that the generalized U-matrix visualization of a topographic map (second DBS module) is able to visualize both compact and connected structures. In terms of the preservation of high-dimensional structures, it is a suitable approach for visualizing the BPEs, FPEs and discontinuities in a data set. However, as shown in Fig. 5.6 in chapter 5, this visualization technique has certain limitations. If additional gaps with intruding points are added by the projection method, then the generalized U-matrix is not able to distinguish identical clusters from distinct ones. To the author's knowledge, the only visualization that shows whether clusters have been disrupted uses a linear gray-scale approach based on a holistic solution called the proximity measure [Aupetit, 2007]. In the two-dimensional projected space, Voronoi cells are filled with brighter or darker luminances depending on their high-dimensional distances D to a reference point. "Points with bright cells are connected in the original space" [Aupetit, 2007, p. 17]. However, cluster disruption can only be successfully visualized when the user selects the correct reference point. To estimate the correct reference point for a projected space, additional visualizations of other measures, as introduced in this paper, must be used. Consequently, this process is both time-consuming and challenging and requires user supervision.

Many quality criteria exist for evaluating the visualization of a scatter plot. Chapter 6 addressed the question of whether the currently existing QMs are able to measure structure preservation. By using a generalized, graph-theory-based definition for a neighborhood of points, it is possible to group the QMs based on their semantic characterization. Here, 19 common QMs were reviewed and grouped, and they were compared with regard to their ability to measure the structure preservation of a projection. It is argued here that the QMs that have been presented in the literature have difficulty correctly capturing the discontinuities in high-dimensional data because of their inherent assumptions regarding the underlying high-dimensional structures. This was shown using the Hepta and Chainlink data sets in supplement *A*.

Otherwise, an objective function could be defined using the "best" QM, and it would always be possible to obtain a structure-preserving two-dimensional visualization by optimizing this objective function. In this work, no answer could be found to the question of how the quality of structure preservation can be automatically measured or visualized without prior knowledge.

However, when a prior classification of the data is available, it can be used to evaluate the quality of structure preservation. The structures that should be preserved are defined by such a classification. A QM called the Delaunay classification error (DCE) was developed based on this concept; it allows projections to be ranked and normalized compared with a baseline and also enables statistical testing.

In summary, structure preservation depends on the chosen projection method; however, the task of choosing the correct projection method is challenging because the optimization of an objective function requires the predefinition of the structures to be visualized. The generalized U-matrix is able to visualize the similarities and dissimilarities among high-dimensional data points in a scatter plot of the projected points (BPEs and FPEs), but it is unable to visualize the disruption of clusters, based on which the quality of structure preservation is defined.

The projection method Pswarm

The first module of the DBS framework is called Pswarm. Pswarm is a projection method that does not rely on an objective function. Similarly to SOP, Pswarm uses stigmergy and a swarm of DataBots because swarm techniques are known for their properties of flexibility and robustness [Bonabeau/Meyer, 2001; Şahin, 2004]. However, in contrast to SOP, which uses an ESOM-like grid space, the environment of the DataBots in Pswarm has been redefined based on symmetry considerations [Feynman et al., 2007, pp. 147-153, 745], resulting in the use of polar coordinates on a toroidal hexagonal grid. The combination of symmetry considerations with game theory concepts endows the polar swarm (Pswarm) with a parameter-free annealing process and an automatically selected, data-driven grid size.

The insights presented in chapter 7 demonstrate that Pswarm exhibits both self-organization and swarm intelligence. In the swarm-based techniques presented in the available literature, the swarms used for projection and/or clustering do not take advantage of both concepts (chapter 7.3, Figure 7.4). Moreover, no other reported swarm method exploits game theory or the phenomenon of emergence (as defined in chapter 7, section 3, after [Ultsch, 2007]). Here, the focus is placed on a subfield of dimensionality reduction in which projection methods are used for visualizing high-dimensional data in a two-dimensional space, as opposed to manifold learning methods, which are designed only to find manifolds, not to compress them into two-dimensional space [Venna et al., 2010, p. 2].

Of the methods of projecting high-dimensional data into two-dimensional space, two stand out: Neighborhood Retrieval Visualizer (NeRV) [Venna et al., 2010] and ESOM [Ultsch, 1999]. NeRV optimizes the objective function that quantifies the cost, defined as information retrieval, with the goal of visualizing the similarity relationships between data points. NeRV attempts to achieve a faithful representation of the data in two dimensions by minimizing the BPE and FPE. The cost is a tradeoff between the FPE and BPE[78], which is defined by the parameter λ. ESOM is an unsupervised neural learning algorithm and can be used as a projection method if a large number of neurons is specified. ESOM remains a reference tool for two-dimensional visualization [Lee/Verleysen, 2007, p. 244]. Instead of an objective function, ESOM uses the powerful concept of emergence [Ultsch, 2007] in addition to the 3D visualization technique of [Thrun et al., 2016a], which is based on the U-matrix [Ultsch, 2003a]. Both NeRV and ESOM are state-of-the-art methods for the visualization of high-dimensional data.

Pswarm was compared with the following common projection methods: principal component analysis (PCA), curvilinear component analysis (CCA), t-distributed stochastic neighbor embedding (t-SNE), ESOM, NeRV and the multidimensional scaling (MDS) technique of Sammon mapping. Five artificial three-dimensional data sets from the FCPS were used to compare these projection methods because of their clearly defined natural clusters. Typically, the QMs

[78] In information retrieval terms, precision and recall.

discussed in the literature indirectly assume that a projection method has a deterministic outcome. A problem that has, thus far, remained undiscussed is the stochastic outcomes of some common projection methods, such as t-SNE and CCA. Therefore, the DCEs were calculated for 100 trials per projection method and data set. Thus, the outcomes of the projection methods could be statistically compared. To enable an unbiased comparison, the DCE requires a prior classification that defines the structures in a data set. However, as discussed by [Färber et al., 2010], natural data sets may have more than one useful classification, depending on the context and the algorithm applied, because no universal definition of a cluster exists [Hennig, 2015b, p. 705]. Therefore, the evaluation of different projections methods by DCE only makes sense on artificial data sets with predefined natural clusters (see chapter 9). This is a major limitation of the DCE QM.

It was shown that the two-dimensional projections generated by Pswarm are comparable to those produced by the state-of-the-art methods NeRV and ESOM. To the author's knowledge, every projection method considered here (except ESOM and SOP) optimizes an objective function, which may lead to the disadvantages discussed above. Moreover, some projection methods, such as ESOM and CCA, use a sophisticated annealing scheme that may be sensitive to one or more parameters or have one or more sensitive parameters themselves (e.g., λ in NeRV). Examples are given in chapter 10.2, Tab. 10.1. In contrast to NeRV, Pswarm is not sensitive to any parameter or, as in the case of ESOM, to an annealing scheme and lattice size. It was shown that a projection with minimal BPE and FPE values does not necessarily achieve structure preservation. In the case of NeRV, it was shown that this algorithm is sensitive to its random initialization process (chapter 5, Fig. 5.6, and chapter 10). Venna et al. also proposed an alternative PCA-based initialization [Venna et al., 2010, p. 459], which in itself makes prior assumptions regarding the relevant structures of the high-dimensional data[79], as illustrated by the baseline used to analyze the DCE results (see chapter 10.2 Figure 10.5). Unlike NeRV, Pswarm does not visualize cluster structures if such structures do not exist in the data, as in the case of the Golf Ball data set (or the various continuous data sets presented in supplement D); moreover, because Pswarm is a swarm-based technique, it is more robust to the random initialization process (e.g., the DBS visualization of the leukemia data set in chapter 11, Figure 10.1).

In the third section of chapter 10, the SOP algorithm is emphasized because it is another method based on a swarm of DataBots, as introduced in [Herrmann, 2009]. In [Herrmann, 2011], it was shown that SOP is nearly as good as or even better than the best of its carefully parameterized competitor methods, namely, CCA, t-SNE and ESOM, in terms of the 1-nearest-neighbor classification accuracy and the specially formulated dispersion measure of [Herrmann, 2011, p. 101]. It was also noted that these methods resulted in severe misrepresentations of the structures for several data sets, which was not the case for SOP (see also the scatter plots in section A2 of [Herrmann, 2011, pp. 158-161]).

Notably, the annealing process of the SOP algorithm is not truly self-adaptive; rather, it is parameterized, which can lead to severe errors in the projections. In the best case, the choice of the lattice size and, therefore, the maximal neighborhood radius as well as the choices of the two magic numbers (the jumping DataBots threshold and the maximum number of iterations) in the SOP algorithm have only a minor effect on the visualization of the high-dimensional

[79] PCA maximizes the variance.

structures (as in the cases of the Atom and Chainlink data sets). In the worst case, as for the EngyTime or Iris data set, all structures are prevented from emerging. Moreover, in the case of EngyTime, it was shown that when there is no restriction ensuring that no more than one DataBot can occupy each lattice position, the information about the high-dimensional structure is lost. Unlike the dispersion measure and 1-nearest-neighbor classification approach of Herrmann, in comparison with SOP and based on a topographic map of projected points, the visualizations presented in this work illustrate important improvements achieved by Pswarm, which are described in the last section of chapter 10.

Several examples were presented to demonstrate that the process leading to emergence is disrupted in the SOP algorithm. Other swarms do not exhibit self-organization but instead rely on the optimization of an objective function, which makes emergence impossible. To the author's knowledge, the game theory approach to behavior-based systems remains undiscussed in the available literature on artificial intelligence in data science. The naturally clustered Wine, Swiss Banknotes and Iris data sets all illustrate the importance of consistent and appropriate definitions of the neighborhoods, scents, grid or lattice size and data-driven annealing scheme used for clustering and projection. If these definitions are oblique, as is the case for SOP, then the self-organization of the DataBots is disrupted. The ultimate disruption of the process leading to emergence may be minor (Swiss Banknotes) or major (Wine, Iris), depending on the data set and the specific trial. For the Wine data set, Pswarm gains an advantage because of the ability to choose different a distance whereas the SOP algorithm does not. [Herrmann, 2011, p. 65]. Pswarm allows the user to define a non-metric distance method without any restrictions.

The correct selection of the parameters for the annealing scheme requires an experienced user. For example, it was shown that with the default settings, the ESOM algorithm sometimes projects three, instead of two, clusters for the Atom data set (chapter 5, Fig. 5.6). To further substantiate this argument, additional ESOM projections generated with the default parameters are presented in Supplement E. For example, it is necessary to change the lattice type from toroidal (default) to planar to achieve a correct projection of the Wing Nut data set. If the default parameters are not changed, the structures are very difficult to see. Disruption of the clusters can be seen in the ESOM/U-matrix visualizations of the Iris, Wine, and Swiss Banknotes data sets, in which one or more of the other eight parameters play an important role (see supplement C for these U-matrix visualizations).

Thus, it is argued here that the ESOM/U-matrix projections of the EngyTime, Wing Nut, Iris, Wine and Swiss Banknotes data sets may be misleading because the toroidal ESOM projections are computed without accounting for symmetry considerations, which results in unwanted boundary effects. For example, the maximal radius is set to the diagonal length[80] $\sqrt{L^2 + C^2}$ instead of $L / 2$, which leads to overlapping of the neighborhoods if the neighborhood function is defined as Gaussian. Several examples illustrate that the uniform distribution used in the ESOM and SOP algorithms has no advantages; however, it may have some disadvantages. The attempt to distribute the projected points uniformly on the lattice is useful only if a visualization method is able to reveal the high-dimensional structures of the data. For this reason, the U-matrix visualization [Ultsch, 2003a] is mandatory for ESOM projections. In other cases, uni-

[80] L is the number of lines in the grid, and C is the number of columns.

formly distributed projected points do not lead to new knowledge about the data set. By contrast, for the generalized U-matrix, there is no requirement for the projected points to be uniformly distributed. Consequently, Pswarm outperforms ESOM on density-based data sets such as EngyTime.

Being a swarm-based method, DBS suffers from the disadvantage of high computational costs. When the number of DataBots[81] is greater than 4000, the use of Pswarm is impractical because of the long calculation time. Further research is necessary on the application of game theory as the foundation for a data-driven annealing scheme. At this point, it can be proven only that a weak Nash equilibrium will be found [Nash, 1951], which may be the reason for the high variance observed in the DCE results (chapter 10, section 2). Only with DBS clustering can the variance of the results be noticeably improved. The structures of 14 of the investigated data sets were preserved using Pswarm (chapters 10 and 11).

The main drawbacks of the proposed approach are as follows. If no prior classification is available for a data set, then the use of DCE measure is limited. Thus, it is very difficult to evaluate whether Pswarm and the generalized U-matrix produce a structure-preserving visualization or whether the clusters are disrupted in the visualization. Additionally, the variance of the results remains high: because it is a stochastic projection method, two different trials of Pswarm could yield different visualizations of the same data set. If the number of clusters is known beforehand, *deep swarming* may be able to solve this problem, as the Tetragonula data set demonstrated[82]. Moreover, it should be possible for the swarm to iteratively add new data points during or after the algorithm following a well-defined process**.** At present, the Pswarm algorithm is unable to do this. Briefly, it was demonstrated in sections 2 and 3 of chapter 10 that finding the correct grid or lattice size and annealing scheme for ESOM/SOP may be challenging. It should be emphasized that unlike SOP and, especially, ESOM (see supplement C and E), Pswarm is able to successfully project density-based data sets. The comparison between Pswarm and the other common projection methods with their default parameter settings resulted in two major findings. First, the state-of-the-art methods ESOM and NeRV do not outperform Pswarm, and second, Pswarm has one important advantage, namely, that it is parameter-free. However, if prior knowledge of the data set to be analyzed is available, then a projection method that is appropriately chosen with regard to the structures that should be preserved will always outperform Pswarm. Furthermore, other projection methods may also outperform Pswarm if their settings are carefully selected by an experienced user. In summary, to the author's knowledge, Pswarm is the first swarm-based technique to show emergent properties while simultaneously combining swarm intelligence, self-organization and game theory.

Knowledge discovery with DBS

Up to this point, mainly artificial data sets have been used to assess the capabilities of DBS. In the case of natural data sets, only the prior classifications were considered. However, the introduction of a new clustering method is necessary only if it is useful. Therefore, three complex real world data sets were first analyzed using DBS to confirm its ability to reproduce known knowledge. Subsequently, two high-dimensional data sets were clustered using DBS to obtain

[81] Which is equal to the number of high-dimensional data points.
[82] for details see next section or chapter 11, section 3.

new knowledge. The silhouette plots and the heatmaps, which showed small intracluster distances and large intercluster distances, indicated that the clustering results for all five data sets were valid.

The visualization and connected clustering of the high-dimensional[83] leukemia data set, which contains clearly defined natural clusters (see chapter 3), successfully reproduced the diagnoses of three types of leukemia: acute myeloid leukemia (AML), acute promyelocytic leukemia (APL) and chronic lymphocytic leukemia (CLL). Aside from two outliers (patients), the prior classification of healthy patients and patients diagnosed with the three leukemia subtypes was reproduced by the DBS clustering and visualization. The two outlier patients may be misdiagnosed; however, a future publication will address this diagnostic problem. Chapter 6 showed that aside from ESOM, no other common projection method was able to visualize the predefined cluster structure of this data set. Similarly, in chapter 3, it was demonstrated that common clustering algorithms failed to correctly cluster the leukemia data set, with the exception of the Ward algorithm, which was not able to find the two outliers.

When the dynamic time-warping distance definition was applied on a data set consisting of the gross domestic product (GDP) per capita in 190 countries for the years 1970–2010, two clusters and one outlier were found using DBS. Upon the application of Classification and Regression Tree (CART) analysis, it was found that the two clusters could be explained as being distinguished by the influence of the tragic event of planes crashing into the World Trade Center in 2001.

DBS found 10 clusters in the Tetragonula data set, as verified by the heatmap and silhouette plot. When the largest within-cluster gap, the cluster separation, and the average within-cluster dissimilarity of [Hennig, 2014] were calculated, the resulting values were the minima reported in [Hennig, 2014], presented there in Fig. 4. The 10 identified clusters strongly depended on the locations of the bees (chapter 11, Figure 11.8). Additionally, the application of DBS to this data set illustrated the possibility of using multiple swarms by means of parallel computing, for which the term *deep swarming* (see [Ultsch, 2016b]) is introduced here in analogy to deep learning [Goodfellow et al., 2016]. Here, deep swarming was applied with a DCE-based objective function, but it can also be applied in combination with any arbitrary objective function.

For the hydrology data set, the daily courses were analyzed. After preprocessing, DBS identified five distinct clusters (chapter 12, Figure 11.4), which were verified by the heatmap and silhouette plot. The rules extracted from a CART decision tree were applied to the clustering of this data set and found to result in the misclassification of 0.9% of the points (chapter 12, Figure 12.6). Five different water quality states in terms of nitrate concentration and electrical conductivity were identified based on a semantic characterization of these clusters (chapter 12, Figure 12.7). The extracted rules enable the prediction of future nitrate and electrical conductivity conditions.

For the pain gene data set, focus was placed on the task of clustering the pain genes. The distances between genes were defined based on the inverse document frequency (idf) [Sparck Jones, 1972] and the information available in the Gene Ontology (GO) database. The DBS clustering resulted in eight clusters (Figure 12.9). Five clusters reproduced the previously known functions of the pain genes (Tab 12.2), as described in section 12.2.1. Outliers were

[83] Containing 7747 variables.

found in two clusters, and one cluster yielded new discoveries regarding the functions of pain genes (Tab 12.2, C5). This cluster was characterized by the downregulation of metabolic processes and the upregulation of the creatine metabolic process.

> *"The experience from many knowledge discovery tasks ([Behnisch/Ultsch, 2009; Kupas et al., 2004; Lötsch/Ultsch, 2013; Mörchen et al., 2005]) is that about 80% of clusters coincide with known processes. Typically about 10% may be attributed to erroneous data, while the remaining 10% may generate entirely new knowledge" [Behnisch/Ultsch, 2015, p. 68].*

This experience is consistent with the findings obtained in the above examples. Two domain experts found the results presented above to be valid and useful.