1 Introduction

We live in a time when information is cheaply available and saved as data nearly everywhere. The amount of generated data is growing exponentially. By the end of the year 2016 alone, 9000 exabytes of data will have been generated, equal to 9 trillion gigabytes or the capacity of 360 billion Blu-ray Discs [Schiele, 2016]. The goal of the interdisciplinary field of data science is to extract knowledge from these data with the help of statistics, machine learning or data mining. Unlike in physics, a data scientist hardly ever starts with a hypothesis; he also is not interested in the source of the data or how they were collected. The data must be mined to gain knowledge through the identification of consistent patterns, and this is usually a very trying task.

Among the various available methods of analyzing data, the focal point of this work is cluster analysis. In contrast to common approaches, the goal here is not merely to group similar information but also to explain why the grouping of information in a certain context is valid, non-trivial and useful. Only then will the clustering of data be helpful to a domain expert. Cluster analysis "is a discipline on the intersection of different fields and can be viewed from different angles, which may be sometimes confusing because different perspectives may contradict each other" [Mirkin, 2005, p. 33]. From the statistical perspective, some assumption regarding the underlying model is required, and data clusters are viewed as probability distributions whose properties can be estimated from the data themselves [Mirkin, 2005, pp. 33-34]. "A trouble with this approach is that in most cases clustering is applied to phenomena of which nothing is known" [Mirkin, 2005, p. 34]. Here, cluster analysis is regarded as the process of generating a classification based on empirical data in a situation in which clear theoretical concepts and definitions are absent and the patterns and laws governing the situation are unknown (see [Mirkin, 2005, p. 36]). The concept of every application (available as open-source code in the R language [R Development Core Team, 2008]) used throughout this thesis is based on this idea.

The goal of this work is to provide an open-source framework for cluster analysis that is founded on a swarm-based projection method and uses a human-understandable visualization approach based on a topographic map of high-dimensional data structures, with the option of 3D printing (see [Thrun et al., 2016a]). This framework should be sufficiently stable while remaining adaptive and exhibiting sufficient plasticity to permit the creation of clusters of various shapes. It should include only a very few non-sensitive parameters that can be visually deduced by a non-professional data miner without any need to understand the theory behind them.

To achieve this goal, expertise on various topics from various areas of research will be required. It is the author's experience that experts in different fields rarely share or exchange practical approaches, and almost nobody is interested in providing and willing to provide easily available and human-understandable solutions to domain experts.

Here, the main hope is to be able to provide reproducible cluster analysis solutions for non-professional data miners and to deliver human-understandable concepts of high-dimensional data structures that are simultaneously able to be processed by machines. In the context of the Databionic swarm (DBS) approach, the author attempts to build, use and explain connections

M. C. Thrun, *Projection-Based Clustering through Self-Organization and Swarm Intelligence*, https://doi.org/10.1007/978-3-658-20540-9_1

among various fields of research; to be precise, the author will illustrate connections between cluster analysis [Hennig et al., 2015; Jain/Dubes, 1988], the imitation of collective behavior [Beni/Wang, 1993; Bonabeau et al., 1999; Reynolds, 1987], the visualization of information [Venna et al., 2010] and its evaluation, machine learning applications [Herrmann/Ultsch, 2008c], game theory [Nash, 1951], symmetry considerations in physics [Feynman et al., 2007, pp. 147-153, 745] and emergence [Ultsch, 2007]. Undoubtedly, making connections between different schools of thought sometimes requires simplifications. For example, with regard to the collective behavior of bees, the fact that bees have a queen who influences their behavior remains unaddressed in this work. Such simplifications are necessary for analytical modeling and applications of cluster analysis.

Chapter 2 addresses most of the necessary definitions and lays the groundwork for all of the mathematical notation used throughout the thesis. The literature reviewed in chapter 3 shows how common clustering methods tend to implicitly assume the patterns or structures sought in data. The reviewed clustering methods are grouped based on their definitions of generalized neighborhoods.

Chapter 4 introduces and classifies common methods of projecting high-dimensional data into two dimensions. Such projections are necessary to cope with the pitfalls of higher dimensions (see, e.g., [Bouveyron/Brunet-Saumard, 2014, pp. 55-57; Verleysen et al., 2003]). Two- or three-dimensional projections will always result in errors; however, gaining a spatial understanding of more than three dimensions is typically an excessively complex task for humans.

Chapter 5 presents examples to depict the typical errors encountered and describes efforts to manage these errors by means of the U-matrix visualization approach [Ultsch, 2003a]. By contrast, chapter 6 demonstrates a more stringent mathematical approach based on quality measures (QMs) presented in the literature. The evaluation of 19 QMs yields a grouping of the QMs based on their implied characterization of structures of high-dimensional data using the definition of neighborhoods introduced in this thesis. Consequently, it is not possible to generalize any of the QMs. If it were possible, the corresponding optimization approaches would not imply any prior assumptions about the structures of high-dimensional data and, consequently, would outperform any other projection methods.

Chapter 7 discusses a nature-inspired and behavior-based system of data science with the goal of using emergence, instead of the optimization of an objective function, for data visualization and clustering.

Building on the insights gained in chapter 7, chapter 8 introduces the DBS concept. Because it relies on the self-organization of data and emergence, DBS does not imply any particular structure that is sought in data. In the context of the projection, visualization and clustering of artificial or high-dimensional data, chapters 10-12 compare DBS with various common methods and apply the DBS framework both to reproduce known insights and to gain new knowledge about various types of data, e.g., multivariate time series or genetic data.

Readers may skip certain chapters depending on their interests. However, the contents of some chapters are based on insights from previous chapters, as indicated by arrows in Figure 1.1, which outlines the organization of this work. Please note, that due to technical limitations the figures and equations are numbered chapter wise.

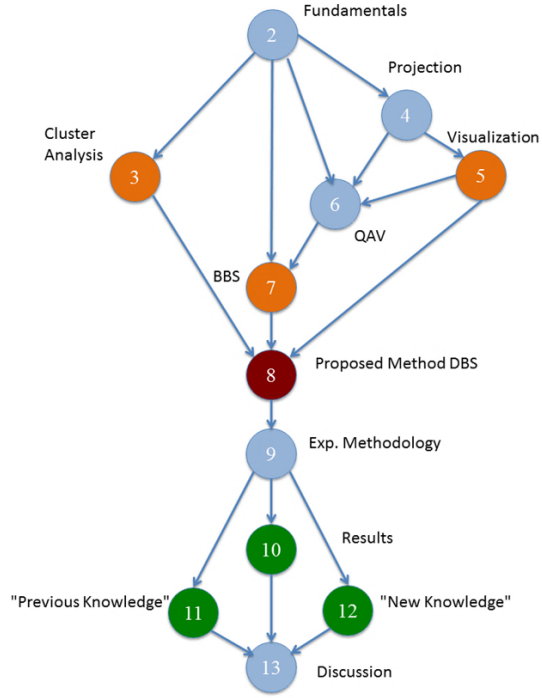

Figure 1.1: Dependency graph of the chapters. BBS: behavior based systems; QAV: Quality Assessments of Visualizations; DBS: Databionic swarm. The underlying concept of DBS is based on insights from chapters 3, 5 and 7 (orange). The evaluation of DBS is performed in three steps (green): general validation in chapter 10, the reproduction of known knowledge in chapter 11, and the generation of new knowledge, as validated by domain experts, in chapter 12.