On Topological Data Mining

Andreas Holzinger[1,2]

[1] Research Unit Human-Computer Interaction, Institute for Medical Informatics,
Statistics & Documentation, Medical University Graz, Austria
a.holzinger@hci4all.at

[2] Institute for Information Systems and Computer Media, Graz University of
Technology, Austria
a.holzinger@tugraz.at

Abstract. Humans are very good at pattern recognition in dimensions of ≤ 3. However, most of data, e.g. in the biomedical domain, is in dimensions much higher than 3, which makes manual analyses awkward, sometimes practically impossible. Actually, mapping higher dimensional data into lower dimensions is a major task in Human–Computer Interaction and Interactive Data Visualization, and a concerted effort including recent advances in computational topology may contribute to make sense of such data. Topology has its roots in the works of Euler and Gauss, however, for a long time was part of theoretical mathematics. Within the last ten years computational topology rapidly gains much interest amongst computer scientists. Topology is basically the study of abstract shapes and spaces and mappings between them. It originated from the study of geometry and set theory. Topological methods can be applied to data represented by point clouds, that is, finite subsets of the n-dimensional Euclidean space. We can think of the input as a sample of some unknown space which one wishes to reconstruct and understand, and we must distinguish between the ambient (embedding) dimension n, and the intrinsic dimension of the data. Whilst n is usually high, the intrinsic dimension, being of primary interest, is typically small. Therefore, knowing the intrinsic dimensionality of data can be seen as one first step towards understanding its structure. Consequently, applying topological techniques to data mining and knowledge discovery is a hot and promising future research area.

Keywords: Computational Topology, Data Mining, Topological Data Mining, Topological Text Mining, Graph-based Text Mining.

1 Introduction and Motivation

Medicine, Biology and health care of today is challenged with complex, high-dimensional, heterogenous, noisy, and weakly structured data sets from various sources [1]. Within such data, relevant *structural* patterns and/or *temporal* patterns ("knowledge") are often hidden, difficult to extract, hence not accessible to a biomedical expert.

A. Holzinger, I. Jurisica (Eds.): Knowledge Discovery and Data Mining, LNCS 8401, pp. 331–356, 2014.

Consequently, a grand challenge is to interactively discover unknown patterns within such large data sets. Computational geometry and algebraic topology may be of great help here [2] embedded in understanding large and complex data sets. Vin de Silva (2004) [3] in his research statement brought the basic idea straight to the point: Let M be a topological space, known as the hidden parameter space; let $\mathbb{R}^D$ be an Euclidean space, defined as observation space, and let $f : M \to \mathbb{R}^D$ be a continuous embedding; $X \subset M$, be a finite set of data points, and $Y = f(X) \subset \mathbb{R}^D$ be the image of these points under the mapping f. Consequently, we may refer to X as the hidden data, and Y as the observed data. The central question then is: Suppose M , f and X are unknown, but Y is known: can we identify M?

This paper of course can only be a scratch on the sheer endless surface, however, the main intention is in motivation and stimulation of further research and to provide a rough guide to the concepts of topology for the non-mathematician, with open eyes on applicability in knowledge discovery and data mining. The paper is organized as follows: In section 2 some key terms are explained to ensure a common and mutual understanding. It is always good to know a bit about who was working in the past in these areas and who are the current leading researchers, consequently in section 3 a very short look on the past is given, followed by section 4 with a brief look on the present. Section 5 provides a nutshell-like overview on the basics of topology, introducing the concepts of point clouds and spaces, manifolds, simplicial complexes and the alpha complex. In chapter 6 a short view on the state-of-the-art in topological data mining, and topological data analysis, respectively, is given; followed by topological text mining in chapter 7. A few software packages are listed in chapter 8, and a few open problems are described in chapter 9. The paper finishes with a section on future challenges and a conclusion with a one-sentence outlook into the future.

2 Glossary and Key Terms

Algebraic Topology: the older name was combinatorial topology, is the field of algebra concerned with computations of homologies and homotopies and other algebraic models in topological spaces [4]. Note: *Geometric topology* is the study of manifolds and embeddings of manifolds.

Alpha Shapes: is a family of piecewise linear simple curves in the Euclidean plane associated with the shape of a finite set of points [5]; i.e. α-shapes are a generalization of the convex hull of a point set: Let $\mathbf{S}$ be a finite set in $\mathbb{R}^3$ and α a real number $0 \leq \alpha \leq \infty$; the u-shape of $\mathbf{S}$ is a polytope that is neither necessarily convex nor necessarily connected. For $\alpha \to \infty$ the α-shape is identical to the convex hull of $\mathbf{S}$ [6]. α-shapes are important e.g. in the study of protein-related interactions [7].

Betti Number: can be used to distinguish topological spaces based on the connectivity of n-dimensional simplicial complexes: In dimension k, the rank of the k-th homology group is denoted β_k, useful e.g. in content-based image retrieval in the presence of noisy shapes, because Betti numbers can be used as shape descriptor admitting dissimilarity distances stable under continuous shape deformations [8].

Computational geometry: A field concerned with algorithms that can be defined in terms of geometry (line segments, polyhedra, etc.) [9].

Contour: ia a connected component of a level set $h - 1(c)$ of a Morse function $h : M \rightarrow R$ defined on a manifold M.

Delaunay triangulation: Given a set of points in a plane $P = p_1, ..., p_n$, a Delaunay triangulation separates the set into triangles with $p's \in P$ as their corners, such that no circumcircle of any triangle contains any other point in its interior [10].

Euler characteristic χ: is an integer associated to a manifold, e.g. χ of a surface is given by the number of faces minus edges plus vertices [11].

Gromov-Norm: is an invariant associated with the homology of a topological space that measures how many simplices are needed to represent a given homology class.

Hausdorff-Space: is a topologically separated space. Let x and y be two distinct points in a topological space X. Let U be the neighbourhood of x and V be the neighborhood of y. x and y are said to be separable if $U \cap V = \emptyset$. Then X is a Hausdorff-Space if every possible pair of points x, y it contains are separable. A Hausdorff space is defined by the property that every two distinct points have disjoint neighborhoods.

Homomorphism: is a function that preserve the operators associated with the specified structure.

Homological algebra: is the study of homology and cohomology of manifolds. Homological algebra is a grand generalization of linear algebra.

Homotopy: Given two maps $f, g : X \rightarrow Y$ of topological spaces, f and g are homotopic, $f \simeq g$, if there is a continuous map $H : X \times [0, 1] \rightarrow Y$ so that $H(x, 0) = f(x)$ and $H(x, 1) = g(x)$ for all $x \in X$ [12].

Homology: Homology and cohomology are algebraic objects associated to a manifold, which give one measure of the number of holes of the object. Computation of the homology groups of topological spaces is a central topic in classic algebraic topology [13]; if the simplicial complex is small, the homology group computations can be done manually; to solve such problems generally a classic algorithm exists, see: [14].

Isometry: is a mapping of metric spaces which preserves the metric.

Metric space: A space in which a distance measure between pairs of elements (points) exists. Note: a metric is a distance function on a space or set; an assignment of distance to every unordered pair of points that satisfies the triangle inequality.

Manifold: is a fundamental mathematical object which locally resembles a line, a plane, or space.

Persistent Homology: Persistent homology is an algebraic tool for measuring topological features of shapes and functions. It casts the multi-scale organization we frequently observe in nature into a mathematical formalism. Here we give a record of the short history of persistent homology and present its basic concepts. Besides the mathematics we focus on algorithms and mention the various connections to applications, including to biomolecules, biological networks, data analysis, and geometric modeling [15]. The concept of persistence emerged independently in the work of Frosini, Ferri et al., and in the thesis of Robins at Boulder, Colorado, and within the biogeometry project of Edelsbrunner at Duke, North Carolina.

Point clouds: are finite sets equipped with a family of *proximity* (or *similarity measure*) functions $sim_q \colon S^{q+1} \rightarrow [0,1]$, which measure how "close" or "similar" $(q+1)$-tuples of elements of S are (a value of 0 means totally different objects, while 1 corresponds to essentially equivalent items).

Reeb graph: is a graph that captures the connectivity of contours; when not having cycles, it is called a contour tree [16]. The Reeb graph is a useful tool in visualizing real-valued data obtained from computational simulations of physical processes [17], [18].

Simplex: is an n-dimensional generalization of the triangle and the tetrahedron: a polytope in n dimensions with $n+1$ vertices.

Simplicial Complex: is made up of simplices, e.g. a simplicial polytope has simplices as faces and a simplicial complex is a collection of simplices pasted together in any reasonable vertex-to-vertex and edge-to-edge arrangement. A graph is a 1-dim simplicial complex.

Space: is generally a set of points $a \in \mathbb{S}$ which satisfy some geometric postulate.

Sphere: is any manifold equivalent (homeomorphic) to the usual round hollow shell in some dimension: a sphere in $n+1$-dimension is called an n-sphere.

Topological Space: is a pair $(\mathbb{X}, \mathbb{T})$ with $\emptyset \in \mathbb{T}$, $\mathbb{X} \in \mathbb{T}$ and a collection of subspaces, so that the union and intersections of subspaces are also in $\mathbb{T}$, in other words, it is a set of points, along with a set of neighbourhoods for each point, that satisfy a set of axioms relating points and neighbourhoods. The definition of a topological space relies only upon set theory and is the most general notion of a mathematical space that allows for the definition of concepts such as continuity, connectedness, and convergence. Other spaces, such as manifolds and metric spaces, are specializations of topological spaces with extra structures or constraints.

Voronoi region: Given a set of points in a plane $p_1, ..., p_n$, a Voronoi diagram erects regions around a point p_i such that all points q within its region are closer to p_i (with regard to some distance measure) than to any other point p_j.

Knowledge Discovery: Exploratory analysis and modeling of data and the organized process of identifying valid, novel, useful and understandable patterns from these data sets.

Minimum Spanning Tree: Given a graph $G = (V, E, \omega)$ with V being the set of vertices, E being the set of edges and ω being the sets of edge weights, a Minimum Spanning tree is the connected acyclic subgraph defined by the subset $E' \subseteq E$ reaching all vertices $v \in V$ with the minimal sum of edge weights possible.

weak/weakly: in mathematics an object is called weak if it is of a generalized kind with fewer properties, and a property holds weakly if it holds in a lesser sense; e.g. a weak solution to an equation might be a discontinuous solution if a straightforward interpretation implies continuity.

3 Topology - The Past

If we want to look into the future, we always should at first look into the past. Topology has its roots in the work on graph theory by Leonhard Euler (1707–1783) [19]. The first book on topology titled “Vorstudien zur Topologie” was published 1848 by Johann Benedict Listing (1808–1882), who emphasized that the term “analysis geometria situs” used by Gottfried Wilhelm Leibniz (1646–1716) was a different geometric concept, hence topology did *not* start before the time of Euler [20]. Listing was very advanced at his time, which can be seen in his 1862 work (see Fig. 1) “Census raeumlicher Complexe” [21]. Significant contributions were made by Carl Friedrich Gauss (1777–1855) and August Ferdinand Moebius (1790–1868), who started with the first steps in set theoretic topology with his 1863 work “Theorie der elementaren Verwandtschaft” [22]. However, topology was not established as own discipline before the formal introduction of *set theory* by Georg Cantor (1845–1918) and Richard Dedekind (1831–1916), the latter was the last student of Gauss.

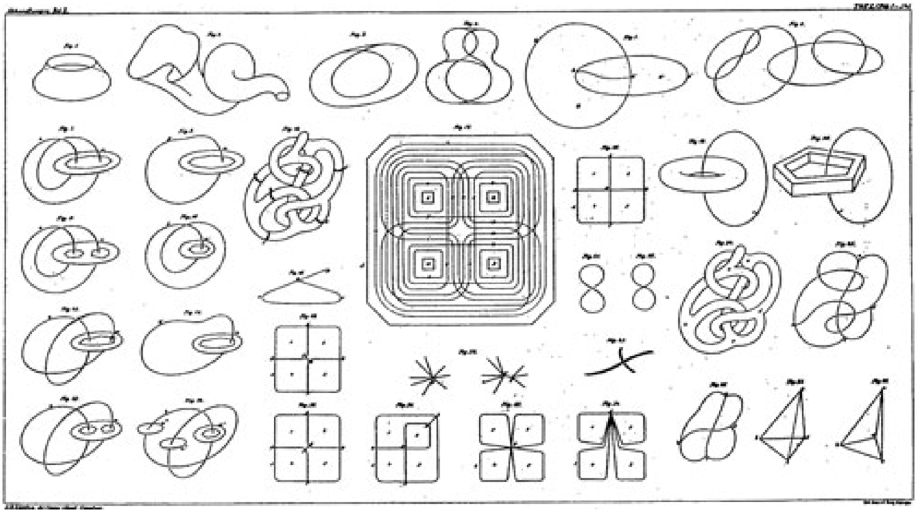

Fig. 1. From the book of Listing (1862)[21]; the image is already in the public domain

Consequently, actual pioneers of (combinatorial, the later algebraic) topology include Henri Poincare (1854–1912), but also Felix Klein (1849–1925), Enrico Betti (1823–1892), Bernhard Riemann (1826–1866) and last but not least Emmi Noether (1882–1935). After these pioneering years the field did not gain much interest, until the discovery of the concept of a *topological space* in 1914 by Felix Hausdorff (1868–1942). In the period after world war I, a collective of mainly French mathematicians pursued these topics amongst others, and they published from 1935 on under a pseudonym called Nicolas Bourbaki. The topics were continued by many meanwhile famous mathematicians, to mention only a few of the "big names": Edwin Evariste Moise (1918–1998), Georges Henri Reeb (1920–1993), Boris Nikolaevich Delaunay (1890–1980), Pavel Sergeyevich Alexandrov (1896–1982) to mention only a few.

According to Blackmore & Peters (2007) [23] the term "computational topology" occurred first in the dissertation of Maentylae in 1983, but there is a journal paper by Tourlakis & Mylopoulus called "Some results in Computational Topology" from 1973 [24] preceded by a conference paper.

4 Computational Topology - The Present

As it is important to look at the past, it is even more important to know some current experts in the field (in alphabetical order, list *not* complete - please forgive shortness and missing names):

Peter Bubenik from the department of mathematics at the Cleveland State University is combining ideas from topology and statistics [25].

Benjamin A Burton from the School of Mathematics and Physics at the University of Queensland, Brisbane, Australia, is the developer of Regina, which is a suite of mathematical software for 3-manifold topologists [26].

Gunnar Carlsson from the Stanford Topology Group, USA, is working in this area for a long time and got famous with his work on "topology and data" and "data has shape" [27].

Tamal K. Dey at the Ohio State University, Columbus, is together with Edelsbrunner and Guha one of the early promoters of computational topology [28].

Nathan Dunfield at the University of Illinois at Urbana-Champaign is working on Topology and geometry of 3-manifolds and related topics and maintaining the CompuTop.org Software Archive (see chapter software) [29].

Herbert Edelsbrunner born in Graz, long time at Duke, North Carolina, is one of the early pioneers in the field and currently at the Institute of Science and Technology Austria in Maria Gugging (near Vienna) [2].

Massimo Ferri , University of Bologna, Italy, was contributing to the concept of persistence, which emerged independently in the work of Cerri, Frosini et al. in Bologna, in the doctoral work of Robins at Boulder, Colorado, and within the biogeometry project of Edelsbrunner at Duke, North Carolina [30].

Robert W. Ghrist at the department of mathematics of the University of Pennsylvania, is particulary working on applied topology in sensor networks [31].

John L. Harer Duke University, Durham, North Carolina, USA, worked a long time together with Edelsbrunner at Duke [2].

Dmitriy Morozov at the Visualization group of the Lawrence Berkeley National Lab, is working on persistent homology [32].

Marian Mrozek is mathematician at the Computer Science Department, Jagiellonian University, Krakw, Poland [33].

Valerio Pascucci at the Center for Extreme Data Analysis and Visualization, University of Utah, applies topological methods to Visualization [34].

Vanessa Robins from the Applied Mathematics department at the Australian National University [35].

Vin de Silva worked with Tenenbaum and Carlsson and is now at Pomona College [36].

Joshua B. Tenenbaum from the Department of Brain and Cognitive Sciences, MIT, Cambridge, Massachusetts, gained much popularity (7097 citations in Google Scholar as of April,18,2014) with the paper in Science on "A Global Geometric Framework for Nonlinear Dimensionality Reduction" [36].

Afra Zomorodian currently working with the D.E. Shaw Group, New York, USA, formerly Department of Computer Science at Dartmouth College, Hanover, New Hamsphire is author of the book "Topology for Computing" [37].

5 Topology in a Nutshell

5.1 Benefits of Topology

Let us start with a thought on our human visual system: We do not see in three spatial dimensions directly, but rather via sequences of planar projections integrated in a manner that is sensed if not comprehended. A newborn does not know what "Google" is, well this is a very abstract example, but the newborn does also not know what an "apple" is. We spend a significant portion of the first decade of our life to learn how to infer three-dimensional spatial data from paired planar projections. Years of practice have tuned a remarkable ability to extract global structure from representations in a strictly lower dimension. Ghrist (2007) [31] starts in the beginning of his paper with summarizing three benefits of topology:

1. It is beneficial to replace a set of data points with a family of simplicial complexes, indexed by a proximity parameter. This converts the data set into global topological objects.
2. It is beneficial to view these topological complexes through the lens of algebraic topology - specifically, via the theory of persistent homology adapted to parameterized families.
3. It is beneficial to encode the persistent homology of a data set in the form of a parameterized version of a Betti number: a barcode.

Algebra and Topology are axiomatic fields, hence would need many definitions, which is impossible to present here, however, before continuing with the main part of this paper, topological data mining, it is necessary to briefly present two fundamental concepts: manifolds and simplicial complexes. Even before, we introduce the primitives of topology: point sets.

5.2 Primitives of Topology: Point Cloud Data Sets

Point cloud data sets (PCD) are the primitives of topology. Consequently, the first question is: "How to get point sets?", or "How to get a graph structure?". Apart from "naturally available" point clouds as discussed below, the answer to this question is not trivial; for some solutions see [38]. In Fig. 2 we see point sets in the plane, resulting from a continuous handwriting input signal given by an input device as

$$X(t) = (x(t), y(t), p(t))^T \tag{1}$$

It contains the coordinates $x(t)$ and $y(t)$ as well as the pressure $p(t)$ of the stylus. After the digitalization process, $X(t)$ is considered as a discrete time series sampled at different points $t \in T$ over time. Let the sampling times be $t_0, t_1, ..., t_n$, satisfying $0 \leq t_0 < t_1 < ... < t_n$. If the time points are equally spaced (*i.e.*, $|\,t_{i+1} - t_i| = \tau$ for all $i = 0, 1, ..., n-1$, $\tau > 0$ some constant), we call the input signal *regularly* sampled.

Let $d(X(t_i), X(t_{i+1})) = \left((x(t_{i+1}) - x(t_i))^2 + (y(t_{i+1}) - y(t_i))^2\right)^{1/2}$ be the Euclidian distance with respect to the coordinates $x(t)$ and $y(t)$. A sampling of the handwriting trajectory satisfying $d(X(t_i), X(t_{i+1})) = \delta$, for some constant $\delta > 0$ and $i = 0, 1, ..., m-1$, is referred as the *equidistant* re-sampling of the time series $X(t)$. We also notice that $t_m \leq t_n$ holds and in general the equidistant re-sampling is not regular (see Fig. 2 on the right).

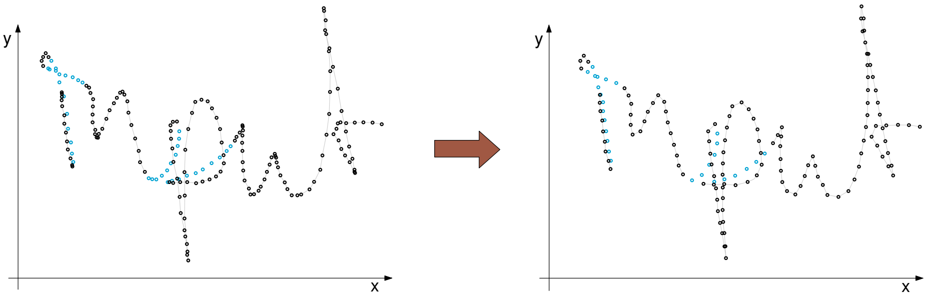

Fig. 2. Point Cloud in $\mathbb{R}^2$ from an handwriting example [39]

Another "natural" source for such point cloud data sets are 3-D Laser scanners (for example the Kinect device). Medical images in nuclear medicine are usually represented in 3D, where a point cloud is a set of points in $\mathbb{R}^3$, whose vertices are characterized by their position and intensity. In dimensions higher than three, point clouds (feature vectors) can be found in the representation of high-dimensional manifolds (see next chapter), where it is usual to work directly with this type of data [40], resulting from protein structures or protein interaction networks [41]. Also in the representation of text data, point clouds appear: Based on the vector space model, which is a standard tool in text mining for a long time [42], a collection of text documents (corpus) can be mapped into a set of points (vectors) in $\mathbb{R}^n$. Each word can also be mapped into vectors, resulting in a very high dimensional vector space. These vectors are the so-called term vectors, with each vector representing a single word. If there, for example, are n keywords extracted from all the documents then each document is mapped to a point (*term vector*) in $\mathbb{R}^n$ with coordinates corresponding to the weights. In this way the whole corpus can be transformed into a point cloud set. Usually, instead of the Euclidean metric, using a specialized similarity (proximity) measure is more convenient. The *cosine similarity measure* is one example which is now a standard tool in text mining, see for example [43]. The cosine of the angle between two vectors (points in the cloud) reflects how "similar" the underlying weighted combinations of keywords are [44].

A set of such primitive points forms a space (see Fig. 3a), and if we have finite sets equipped with proximity or similarity measure functions $sim_q\colon S^{q+1} \rightarrow [0,1]$, which measure how "close" or "similar" $(q+1)$-tuples of elements of S are we have a topological space (see Fig. 3b). A value of 0 means totally different objects, while 1 corresponds to essentially equivalent items. In Fig. 2 we see a good example of a direct source for point clouds in an space which we can easily perceive in $\mathbb{R}^2$. A metric space (see Fig. 3c) has an associated metric (see Fig. 3d the Euclidean distance), enabling to measure distances between points in that space and to define their neighborhoods. Consequently, a metric provides a space with a topology, and a metric space is a topological space.

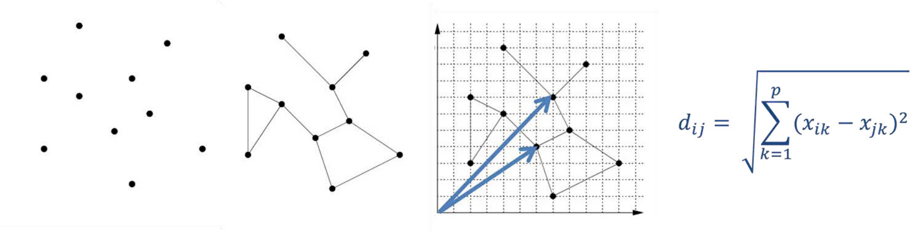

Fig. 3. From left to right: (a) point clouds, (b) point clouds equipped with proximity in a graph structure, (c) points in a metric space $\mathbb{R}$, this is practical because we can (d) measure in this space with the Euclidean distance

5.3 Manifolds

A manifold is a topological space, which is locally homeomorphic (has a continuous function with an inverse function) to a real n-dimensional space (e.g. Euclidean space as in Fig. 3). In other words: X is a d-manifold if every point of X has a neighborhood homeomorphic to $\mathbb{B}^d$; **with boundary** if every point has a neighborhood homeomorphic to $\mathbb{B}$ or $\mathbb{B}^d_+$, in other words it is a topological space which is locally homeomorphic (has a continuous function with an inverse function) to a real n-dimensional space (e.g. Euclidean space) [45].

A topological space may be viewed as an abstraction of a metric space, and similarly, manifolds generalize the connectivity of d-dimensional Euclidean spaces $\mathbb{B}^d$ by being locally similar, but globally different. A d-dimensional chart at $p \in X$ is a homeomorphism $\phi : U \rightarrow \mathbb{R}^d$ onto an open subset of $\mathbb{R}^d$, where U is a neighborhood of p and open is defined using the metric. A d-dimensional manifold (d-manifold) is a topological space X with a d-dimensional chart at every point $x \in X$.

The circle or 1-sphere S^1 in Fig. 4(a) is a 1-manifold as every point has a neighborhood homeomorphic to an open interval in $\mathbb{R}^1$. All neighborhoods on the 2-sphere S^2 in Fig. 4(b) are homeomorphic to open disks, so S^2 is a 2-manifold, also called a surface. The boundary ∂X of a d-manifold X is the set of points in X with neighborhoods homeomorphic to $H^d = x \in \mathbb{R}^d | x_1 \geq 0$. If the boundary is nonempty, we say X is a manifold with boundary. The boundary of a d-manifold with boundary is always a $(d-1)$-manifold without boundary. Figure 4(c) displays a torus with boundary, the boundary being two circles [46].

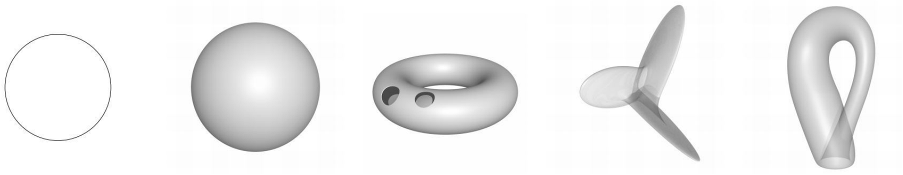

Fig. 4. Manifolds. From left to right: (a) a circle S is a 1-manifold; (b) The sphere S^2 is a 2-manifold; (c) The torus is also a 2-manifold with boundaries; (d) A Boys surface is a geometric immersion of the projective plane P^2, thus a non-orientable 2-manifold; (e) The famous Klein bottle is a non-orientable 2-manifold [46].

5.4 Simplicial Complexes

Simplicial complexes are spaces described in a very particular way, the basis is in Homology. The reason is that it is not possible to represent surfaces precisely in a computer system due to limited computational storage. Consequently, surfaces are sampled and represented with triangulations. Such a triangulation is called a simplicial complex, and is a combinatorial space that can represent a space. With such simplicial complexes, the topology of a space from its geometry can be separated, and Zomorodian compares it with the separation of syntax and semantics in logic [46].

Carlsson emphasizes that not every space can be described as a simplicial complex and that each space can be described as a simplicial complex in many different ways and that calculations of homology for simplicial complexes remains the best method for explicit calculation. Because most spaces of interest are either explicitly simplicial complexes or homotopy equivalent to such, it turns out that simplicial calculation is sufficient for most situations.

Let $S = \{x_0, x_1, \ldots, x_n\}$ denote a subset of a Euclidean space $\mathbb{R}^k$. We say S is in general position if it is not contained in any affine hyperplane of $\mathbb{R}^k$ of dimension less than n. When S is in general position, we define the *simplex spanned by* S to be the convex hull $\sigma = \sigma(S)$ of S in $\mathbb{R}^k$. The points x_i are called *vertices*, and the simplices $\sigma(T)$ spanned by non-empty subsets of $T \subseteq S$ are called *faces* of σ By a (finite) *simplicial complex*, we will mean a finite collection $\mathcal{X}$ of simplices in a Euclidean space so that the following conditions hold.

1. For any simplex σ of $\mathcal{X}$, all faces of σ are also contained in $\mathcal{X}$
2. For any two simplices σ and τ of $\mathcal{X}$, the intersection $\sigma \cap \tau$ is a simplex, which is a face of both σ and τ .

Definition 1. *By an abstract simplicial complex X, we will mean a pair $X = (V(X), \Sigma(X))$, where $V(X)$ is a finite set called the vertices of X, and where $\Sigma(X)$ is a subset (called the simplices) of the collection of all non-empty subsets of $V(X)$, satisfying the conditions that if $\sigma \in \Sigma(X)$, and $\emptyset \neq \tau \subseteq \sigma$, then $\tau \in \Sigma(X)$. Simplices consisting of exactly two vertices are called edges.*

Figure 5 shows some examples; for more details and background please refer to the excellent recent notes of Carlsson (2013) [47], and to the books of Zomorodian (2009) [46] and the book of Edelsbrunner & Harer (2010) [2].

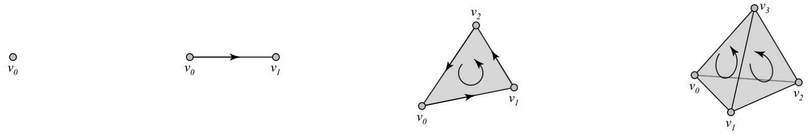

Fig. 5. Oriented k-simplices, $0 \leq k \leq 3$. An oriented simplex induces orientation on its faces, as shown for the edges of the triangle and two faces of the tetrahedron [46].

Topological techniques originated in pure mathematics, but have been adapted to the study and analysis of data during the past two decades. The two most popular topological techniques in the study of data are *homology* and *persistence.* The connectivity of a space is determined by its cycles of different dimensions. These cycles are organized into groups, called homology groups. Given a reasonably explicit description of a space, the homology groups can be computed with linear algebra. Homology groups have a relatively strong discriminative power and a clear meaning, while having low computational cost. In the study of persistent homology the invariants are in the form of persistence diagrams or barcodes [48].

Carlsson [47] defines the persistence vector space as follows:

Definition 2. *Let k be any field. Then by a* persistence vector space *over k, we will mean a family of k-vector spaces $\{V_r\}_{r\in[0,+\infty)}$, together with linear transformations $L_V(r, r') : V_r \to V_{r'}$ whenever $r \leq r'$, so that $L_V(r', r'') \cdot L_V(r, r') = L_V(r, r'')$ for all $r \leq r' \leq r''$. A* linear transformation *f of persistence vector spaces over k from $\{V_r\}$ to $\{W_r\}$ is a family of linear transformations $f_r : V_r \to W_r$, so that for all $r \leq r'$, all the diagrams*

$$\begin{tikzcd}
V_r \arrow[r, "L_V(r,r')"] \arrow[d, "f_r"'] & V_{r'} \arrow[d, "f_{r'}"] \\
W_r \arrow[r, "L_W(r,r')"] & W_{r'}
\end{tikzcd}$$

commute in the sense that

$$f_{r'} \circ L_V(r,r') = L_W(r,r') \circ f_r$$

A linear transformation is an isomorphism *if it admits a two sided inverse. A* sub-persistence vector space *of $\{V_r\}$ is a choice of k-subspaces $U_r \subseteq V_r$, for all $r \in [0,+\infty)$, so that $L_V(r,r')(U_r) \subseteq U_{r'}$ for all $r \leq r'$. If $f : \{V_r\} \to \{W_r\}$ is a linear transformation, then the image of f, denoted by $im(f)$, is the sub-persistence vector space $\{im(f_r)\}$.*

In data mining it is important to extract significant features, and exactly for this, topological methods are useful, since they provide robust and general feature definitions with emphasis on global information.

5.5 Alpha complex

A very important concept which should be mentioned is the so-called **α-complex**: This construction is performed on a metric space X which is a subspace of a metric space Y. Typically Y is a Euclidean space $\mathbb{R}^N$, and most often N is small, i.e. = 2, 3, or 4. For any point $x \in X$, we define the *Voronoi* cell of x, denoted by $V(x)$, by

$$V(x) = \{y \in Y | d(x,y) \leq d(x',y) \text{ for all } x' \in X\}$$

The collection of all Voronoi cells for a finite subset of Euclidean space is called its Voronoi diagram (see Fig. 6).

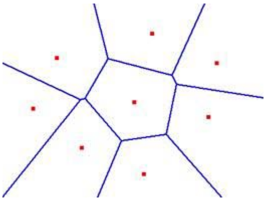

Fig. 6. A picture of part of a Voronoi diagram in $\mathbb{R}^2$ [47], for more details on Voronoi please refer to [49]

For each $x \in X$, we also denote by $B_\epsilon(x)$ the set $\{y \in Y | d(x,y) \leq \epsilon\}$. By the α-cell of $x \in V(x)$ with scale parameter ϵ, we will mean the set $A_\epsilon(x) = B_\epsilon(x) \cap V(x)$. The *$\alpha$-complex with scale parameter* ϵ of a subset $x \in X$, denoted by $\alpha_\epsilon(X)$ will be the abstract simplicial complex with vertex set X, and where the set $\{x_0, \ldots, x_k\}$ spans a k-simplex iff

$$\bigcap_{i=0}^{k} A_\epsilon(x_i) \neq \emptyset$$

An example might look as shown in Figure 7.

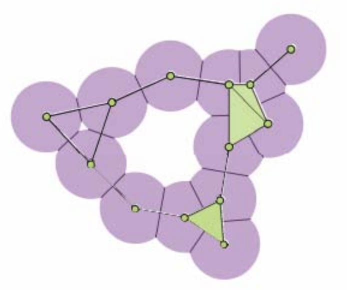

Fig. 7. A typical alpha complex [47], for more details please refer to [6]

6 Topological Data Mining - State-of-the-Art

The term "topological data mining" is still rarely used to date: A Google search as of 30.03.2014 returned only 18 hits in total, a Google Scholar search only 13 hits and a Web of Science search returned *none.* A better known term is "Topological Data Analysis (TDA)", which returns many hits on Google - but on Google Scholar still only 23 results and on the Web of Science 5 hits, however, only three of them are relevant: 1) The editorial on a 2011 special issue in the journal "Inverse Problems" [50] where the editors Charles Epstein, Gunnar Carlsson and Herbert Edelsbrunner emphasize the importance of persistent homology for data analysis;

2) An overview chapter by Afra Zomorodian (2012) [51] in the book "Algorithms and Theory of Computation Handbook, Second Edition, Volume 2: Special Topics and Techniques" by Attalah & Blanton, where he provides on 31 pages a concise overview on Topological Spaces (manifolds, data structures), Topological Invariants (Euler Characteristic, Homotopy), Simplicial Homology, Persistent Homology (and he provides an Algorithm), Morse Theory (Reeb Graph, Morse-Smale Complex), Structures for Point Sets (Geometric Complexes, Persistent Homology);

3) A paper by Blumberg & Mandell (2013) [52] where the authors lay the foundations for an approach to apply the ideas of Michail Gromov on quantitative topology to data analysis. For this purpose they introduce a so-called "contiguity complex", which is a simplicial complex of maps between simplicial complexes defined in terms of the combinatorial notion of contiguity. Moreover, they generalize the Simplicial Approximation Theorem in order to show that the contiguity complex approximates the homotopy type of the mapping space as they subdivide the domain; consequently the authors describe algorithms for approximating the rate of growth of the components of the contiguity complex under subdivision of the domain, which allows to computationally distinguish spaces with isomorphic homology but different homotopy types.

A search with title = "computational topology" resulted also in only 22 hits, the most recent paper, indeed a very good hit: Computational Topology with Regina: Algorithms, Heuristics and Implementations by Burton (2013) [26], where the author documents for the first time in the literature some of the key algorithms, heuristics and implementations that are central to the performance of his software called REGINA; including the simplification heuristics, key choices of data structures and algorithms to alleviate bottlenecks in normal surface enumeration, modern implementations of 3-sphere recognition and connected sum decomposition. The oldest paper is from Tourlaki & Mylopoul (1973) [53], where the authors study topological properties of finite graphs that can be embedded in the n-dimensional integral lattice; they show that two different methods of approximating an n-dimensional closed manifold with boundary by a graph of the type studied in this paper lead to graphs whose corresponding homology groups are isomorphic. This is at the same time the paper with the highest citations, however only 17 (as of April, 18, 2014). A highly cited paper (504 times in the web of science, 1008 in Google Scholar (as of April, 18, 2014) is a survey paper by Kong & Rosenfeld (1989) [54], Digital Topology: Introduction and Survey, however, this is dealing with topological properties of digital images, which is the study of image arrays, not the study of algebraic topology; this must not be mixed up.

7 Topological Text Mining - State-of-the-Art

Maybe the first work on the application of computational topology in text mining was presented at the Computational Topology in *Image* Context conference (CTIC 2012) by [44]. The background is in the vector space model, which is a standard tool in text mining [42]. A collection of text documents (corpus) is mapped into points (=vectors) in $\mathbb{R}^n$. And each word can also be mapped into vectors, resulting in a very high dimensional vector space. These vectors are the so-called term vectors, each vector is representing e.g. a single word. If there are n keywords extracted from all the documents then each document is mapped to a point (*term vector*) in $\mathbb{R}^{\ltimes}$ with coordinates corresponding to the weights. In this way the whole corpus can be transformed into a point cloud set. Usually, instead of the Euclidean metric, using a specialized similarity (proximity) measure is more convenient. The *cosine similarity measure* is one example which is now a standard tool in text mining, see e.g. [43]. Namely, the cosine of the angle between two vectors (points in the cloud) reflects how “similar” the underlying weighted combinations of keywords are. Amongst the many different text mining methods (for a recent overview refer to [55]), a topological approach is very promising, but needs a lot of further research; let us first look on graph-based approaches.

7.1 Graph-Based Approaches for Text Mining

Graph-theoretical approaches for Text Mining emerged from the combination of the fields of data mining and topology, especially graph theory [56]. Graphs are intuitively more informative as example words/phrase representations [57]. Moreover graphs are the best studied data structure in computer science and mathematics and they also have a strong relation with logical languages [56].

Its structure of data is suitable for various fields like biology, chemistry, material science and communication networking [56]. Furthermore, graphs are often used for representing text information in natural language processing [57]. Dependency graphs have been proposed as a representation of syntactic relations between lexical constituents of a sentence. This structure is argued to more closely capture the underlying semantic relationships, such as subject or object of a verb, among those constituents [58].

The beginning of graph-theoretical approaches in the field of data mining was in the middle of the 1990's [56] and there are some pioneering studies such as [59,60,61]. According to [56] there are five theoretical bases of graph-based data mining approaches such as (1) subgraph categories, (2) subgraph isomorphism, (3) graph invariants, (4) mining measures and (5) solution methods. Furthermore, there are five groups of different graph-theoretical approaches for data mining such as (1) greedy search based approach, (2) inductive logic programming based approach, (3) inductive database based approach, (4) mathematical graph theory based approach and (5) kernel function based approach [56].

There remain many unsolved questions about the graph characteristics and the isomorphism complexity [56]. Moreover the main disadvantage of graph-theoretical text mining is the computational complexity of the graph representation. The goal of future research in the field of graph-theoretical approaches for text mining is to develop efficient graph mining algorithms which implement effective search strategies and data structures [57].

Examples in the Biomedical Domain: Graph-based approaches in text mining have many applications from biology and chemistry to internet applications [62]. According to Morales et al [63] graph-based text mining approach combined with an ontology (e.g. the Unified Medical Language System - UMLS) can lead to better automatic summarization results. In [64] a graph-based data mining approach was used to systematically identify frequent co-expression gene clusters. A graph-based approach was used to disambiguate word sense in biomedical documents in Agirre et al. [65]. Liu [66] proposed a supervised learning method for extraction of biomedical events and relations, based directly on subgraph isomorphism of syntactic dependency graphs. The method extended earlier work [67] that required sentence subgraphs to exactly match a training example, and introduced a strategy to enable approximate subgraph matching. These method have resulted in high-precision extraction of biomedical events from the literature.

Discussion: While graph-based approaches have the *disadvantage* of being computationally expensive, they have the following *advantages*:

- It offers a far more expressive document encoding than other methods [57].
- Data which is graph structured widely occurs in different fields such as biology, chemistry, material science and communication networking [56].

7.2 Topological Text Data Mining

Very closely related to graph-based methods are topological data mining methods, due to the fact that for both we need point cloud data sets as input, which can e.g. be achieved by the vector space model, where the tips of the vectors in an arbitrarily high dimensional space can be seen as point data sets [38].

Due to finding meaningful topological patterns greater information depth can be achieved from the same data input [44]. However, with increasing complexity of the data to process also the need to find a scalable shape characteristic is greater [68]. Therefore methods of the mathematical field of topology are used for complex data areas like the biomedical field [68], [48]. Topology as the mathematical study of shapes and spaces that are not rigid [68], pose a lot of possibilities for the application in knowledge discovery and data mining, as topology is the study of connectivity information and it deals with qualitative geometric properties [69].

Functionality: One of the main tasks of applied topology is to find and analyse higher dimensional topological structures in lower dimensional spaces (e.g. point cloud from vector space model [44]). A common way to describe topological spaces is to first create simplicial complexes. A simplicial complex structure on a topological space is an expression of the space as a union of simplices such as points, intervals, triangles, and higher dimensional analogues. Simplicial complexes provide an easy combinatorial way to define certain topological spaces [69]. A simplical complex K is defined as a finite collection of simplices such that $\sigma \in K$ and τ, which is a face of σ, implies $\tau \in K$, and $\sigma, \sigma' \in K$ implies $\sigma \cap \sigma'$ can either be a face of both σ and σ' or empty[70]. One way to create a simplical complex is to examine all subsets of points, and if any subsets of points are close enough, a p-simplex (e.g. line) is added to the complex with those points as vertices. For instance, a Vietoris-Rips complex of diameter ϵ is defined as $VR(\epsilon) = \sigma | diam(\sigma) \leq \epsilon$, where $diam(\epsilon)$ is defined as the largest distance between two points in σ [70]. Figure 8 shows the Vietoris-Rips complex with varying ϵ for four points with coordinates (0,0), (0,1), (2,1), (2,0). A common way a analyse the topological structure is to use persistent homology, which identifies cluster, holes and voids therein. It is assumed that more robust topological structures are the one which persist with increasing ϵ. For detailed information about persistent homology, it is referred to [70].

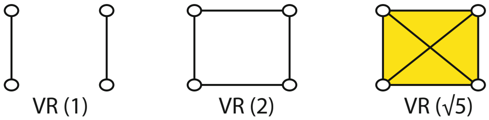

Fig. 8. Vietoris-Rips complex of four points with varying ϵ [70]

Examples in the Biomedical Domain: In [71] a graph-theoretical approach for Text Mining is used to extract relation information between terms in free-text electronic health care records that are semantically or syntactically related. Another field of application is the text analysis of web and social media for detecting influenza-like illnesses [72].

Moreover there can be content-rich relationship networks among biological concepts, genes, proteins and drugs developed with topological text data mining like shown in [73]. According to [74] network medicine describes the clinical application field of topological text mining due to adressing the complexity of human diseases with molecular and phenotypic network maps.

Discussion: A clear *advantage* of topological text mining is that here can be greater information depth achieved through understanding the global structure of the data [44]. The *disadvantages* include

– The Complexity of the graph representation itself is a problem [57].
– There is a performance limitation in handling large datasets in high dimensions [44].

8 Computational Topology: Software

Maybe the most famous algorithm is the one by Delfinado & Edelsbrunner (1995) [75], where the authors present an incremental method for computing the Betti numbers of a topological space represented by a simplicial complex. The algorithm, which has been presented two years earlier at the 9th symposium on computational geometry [76], is an good example of how algorithmic graph techniques can be applied and extended to complexes of dimension higher than one, which was an important step in raising interest for algebraic topology.

Besides from available geometry software, whole software packages in computational topology are rare to date. A good starting point is the CompuTop.org Software Archive maintained by Nathan Dunfield, enlisting prominent packages for computing the homology and cohomology of simplicial complexes and groups; another good source is the CompTop page from Stanford.

Computational Homology Project (CHomP): provides a set of tools for computing the homology of a collection of n-dimensional cubes, with a view towards applied applications in dynamical systems, chaos theory, and pattern characterization, developed by Pawel Pilarczyk and supported by Konsantin Mischaikow, Hiroshi Kokubu, Marian Mrozek, Thomas Gedeon, Jean-Philippe Lessard and Marcio Gameiro.

Dionysus: is a C++ library developed by Dmitriy Morozov for computing persistent homology, distributed together with thin Python bindings. It currently implements persistent homology, vineyards, persistent cohomology, zigzag, alpha shapes, Vietoris-Rips complexes, Čech complexes, circle valued coordinatization, and piecewise linear vineyards.

Homological Algebra Programming (HAP): is a homological algebra library (current version 1.10.15 from November,21, 2013), also for use with GAP with initial focus on computations related to the cohomology of finite and infinite groups [77].

Linbox: is a C++ library with GAP and Maple interfaces for exact, high-performance linear algebra computation with dense, sparse, and structured matrices over the integers and over finite fields [78]. (GAP is a system for computational discrete algebra, with particular emphasis on Computational Group Theory, the current version GAP 4.7.4 released on February, 20, 2014; Maple is the well-known computer algebra system, version 18 released in March 2014).

Mapper: is a software (cf. Patent US20100313157A1) developed by the Stanford Carlsson group (Sexton, Singh, Memoli) [79] for extracting simple descriptions of high dimensional data sets in the form of simplicial complexes, and is based on the idea of partial clustering of the data guided by a set of functions defined on the data. Mapper is the basis of Ayasdi, the company offering the so-called Insight Discovery Platform using Topological Data Analysis (TDA) to allow people to discover insights in data.

Persistent Homology in R (PHOM): is a package by Andrew P Tausz, who graduated in 2013 from the Carlsson Group in Stanford, who also developed JavaPlex. PHPOM is an R package [80] that computes the persistent homology of geometric data sets, to make persistent homology available to the statistics community.

Persistent homology computations (JavaPlex): is a library that implements persistent homology and related techniques from computational and applied topology, enabling extensions for further research projects and approaches. It was developed in 2010/11 by the Stanford CompTop Group to improve JPlex, which is a package for computing persistent homology of finite simplicial complexes, often generated from point cloud data [81].

Regina: is a suite of software for 3-manifold topologists. It focuses on the study of 3-manifold triangulations and normal surfaces. Other highlights of Regina include angle structures, census enumeration, combinatorial recognition of triangulations, and high-level tasks such as 3-sphere recognition and connected sum decomposition. Regina comes with a full graphical user interface, and also offers Python bindings and a low-level C++ programming interface [26].

9 Open Problems

There are many topological algorithms having exponential time complexity and the quest for developing efficient algorithms has started only recently and most of the problems in computational topology still wait for efficient solutions [82]. Some unsolved problems include, for example:

Problem 1. Point cloud data sets or at least distances are the primitives for the application of topological approaches, so unless you do not have direct point data input (e.g. from scanners) the first problem to be solved is in preprocessing, i.e. in transforming data, e.g. natural images into point cloud data sets, which is not a trivial task and poses a lot of problems [38].

Problem 2. Volodin, Kuznetsov and Fomenko (1974) [83] stated the problem of discriminating algorithmically the standard three-dimensional sphere, so an algorithm would be sought that determines whether a simplicial 3-manifold is topological equivalent to $\mathbb{S}^3$, this is a hard problem.

Problem 3. A further open problem is in the design of an algorithm that computes all minimal triangulations for a surface of genus g, or the determination of the minimal size of a triangulation for a triangulable d-manifold; here Vegter provides some pointers to Brehm and Khnel (1987) [84] and Sarkaria (1987) [85].

Problem 4. To date none of our known methods, algorithms and tools scale to the massive amount and dimensionalities of data we are confronted in practice; we need much more research efforts towards making computational topology successful as a general method for data mining and knowledge discovery [46].

Problem 5. A big problem is to compute Reeb graphs for spaces of dimension higher than 3, which would be necessary for knowledge discovery from high-dimensional data [18].

Whilst computational topology has much potential for the analysis of arbitrarily high-dimensional data sets, humans are very good at pattern recognition in dimensions of ≤ 3, this immediately suggests a combination of the "best of the two worlds" towards integrated and interactive solutions [1],[86]. Scientifically, this can be addressed by the HCI-KDD approach: while Human–Computer Interaction (HCI) puts its emphasis on human issues, including perception, cognition, interaction and human intelligence and is tightly connected with Visualization and Interactive Visual Analytics, Knowledge Discovery &Data Mining (KDD) is dealing with computational methodologies, methods, algorithms and tools to discover new, previously insights into data, hence we may speak of supporting human learning with machine learning [87]; the HCI-KDD network of excellence (see www.hci4all.at) is proactively supporting this approach in bringing together people with diverse background but sharing a common goal.

Suppose we were given a million points in 100 dimensions and we wish to recover the topology of the space from which these points were sampled. Currently, none of our tools either scale to these many points or extend to this high a dimension. Yet, we are currently inundated with massive data sets from acquisition devices and computer simulations. Computational topology could provide powerful tools for understanding the structure of this data. However, we need both theoretical results as well as practical algorithms tailored to massive data sets for computational topology to become successful as a general method for data analysis.

10 Conclusion and Future Outlook

Topology is basically the study of shapes, in particular of properties that are preserved when a shape is deformed. Topological techniques originated in pure mathematics in the last 200+ years, and for quite a time it was the playing field of some quirky mathematicians interested in differences between a donut and a dumpling. Meanwhile, topology as mathematical study of shapes and spaces is a mature and established mathematical field, and in the past two decades the principles of topology have been adapted and applied to the study and analysis of data sets, emerging into a very young discipline: *computational* topology. A very popular topological technique related to the study of data sets is homology. The connectivity of a space is determined by its cycles of different dimensions, and these cycles can be organized into groups, so-called homology groups. Given a reasonably description of a space, these homology groups can be computed e.g. by help of linear algebra. Homology groups have a relatively strong discriminative power and a clear meaning at relatively low computational effort. In the study of persistent homology the invariants are in the form of persistence diagrams or so-called barcodes [48].

For knowledge discovery and data mining it is important to visualize and comprehend complex data sets, i.e. to find and extract significant features. Exactly for this reason, topological methods are very useful, since they provide robust and general feature definitions. They emphasize a "global information", although this can lead to problems during parallelization [88]. Rieck et al. (2012) [89] presented a novel method for exploring high-dimensional data sets by coupling topologically-based clustering algorithms with the calculation of topological signatures. Future challenges are in achieving better localization (i.e. assigning a geometrical meaning) of features when using topological signatures. Rieck et al. also suggested that in future research the different ways of creating simplicial complexes should be examined and several metrics for the Rips graph (or neighbourhood graph) should be further investigated. Recently, Morozov (2013) [88] presented a parallel algorithm for merging two trees. They realized that new ideas in this domain will be necessary. Future architectures will have many more cores with non-uniform memory access, hence, an important future research direction is developing data structures that explicitly take asymmetry into account.

A large area of future research is in graph-theoretical approaches for text mining, in particular to develop efficient graph mining algorithms which implement robust and efficient search strategies and data structures [57]. However, there remain much unsolved questions about the graph characteristics and the isomorphism complexity [56], so there are plenty of interesting research lines in the future.

The grand challenge is in the *integration* of methods, algorithms and tools from computational topology into useable and useful solutions for *interactive* knowledge discovery and data mining in high-dimensional and complex data sets.

Acknowledgments. The author is very grateful for fruitful discussions with the hci4all.at Team and with members of the HCI-KDD network of excellence, I thank my Institutes both at Graz University of Technology and the Medical University of Graz, my colleagues and my students for the enjoyable academic freedom, the intellectual environment, and the opportunity to think about crazy ideas following my motto: “Science is to test ideas!”.

References

1. Holzinger, A., Dehmer, M., Jurisica, I.: Knowledge discovery and interactive data mining in bioinformatics state-of-the-art, future challenges and research directions. BMC Bioinformatics 15(suppl. 6), I1 (2014)
2. Edelsbrunner, H., Harer, J.L.: Computational Topology: An Introduction. American Mathematical Society, Providence (2010)
3. De Silva, V.: Geometry and topology of point cloud data sets: a statement of my research interests (2004), `http://pomona.edu`
4. Hatcher, A.: Algebraic Topology. Cambridge University Press, Cambridge (2002)
5. Edelsbrunner, H., Kirkpatrick, D., Seidel, R.: On the shape of a set of points in the plane. IEEE Transactions on Information Theory 29(4), 551–559 (1983)
6. Edelsbrunner, H., Mucke, E.P.: 3-dimensional alpha-shapes. ACM Transactions on Graphics 13(1), 43–72 (1994)
7. Albou, L.P., Schwarz, B., Poch, O., Wurtz, J.M., Moras, D.: Defining and characterizing protein surface using alpha shapes. Proteins-Structure Function and Bioinformatics 76(1), 1–12 (2009)
8. Frosini, P., Landi, C.: Persistent betti numbers for a noise tolerant shape-based approach to image retrieval. Pattern Recognition Letters 34(8), 863–872 (2013)
9. Goodman, J.E., O’Rourke, J.: Handbook of Discrete and Computational Geometry. Chapman and Hall/CRC, Boca Raton (2010)
10. Cignoni, P., Montani, C., Scopigno, R.: Dewall: A fast divide and conquer delaunay triangulation algorithm in ed. Computer-Aided Design 30(5), 333–341 (1998)
11. Bass, H.: Euler characteristics and characters of discrete groups. Inventiones Mathematicae 35(1), 155–196 (1976)
12. Whitehead, G.W.: Elements of homotopy theory. Springer (1978)
13. Alexandroff, P., Hopf, H.: Topologie I. Springer, Berlin (1935)
14. Munkres, J.R.: Elements of algebraic topology, vol. 2. Addison-Wesley, Reading (1984)
15. Edelsbrunner, H., Harer, J.: Persistent Homology - a Survey. Contemporary Mathematics Series, vol. 453, pp. 257–282. Amer Mathematical Soc., Providence (2008)
16. Doraiswamy, H., Natarajan, V.: Efficient algorithms for computing reeb graphs. Computational Geometry 42(67), 606–616 (2009)
17. Edelsbrunner, H., Harer, J., Mascarenhas, A., Pascucci, V., Snoeyink, J.: Time-varying reeb graphs for continuous space-time data. Computational Geometry-Theory and Applications 41(3), 149–166 (2008)
18. Biasotti, S., Giorgi, D., Spagnuolo, M., Falcidieno, B.: Reeb graphs for shape analysis and applications. Theoretical Computer Science 392(13), 5–22 (2008)
19. Euler, L.: Solutio problematis ad geometriam situs pertinentis. Commentarii Academiae Scientiarum Petropolitanae 8(1741), 128–140
20. Listing, J.B.: Vorstudien zur Topologie. Vandenhoeck und Ruprecht, Goettingen (1848)
21. Listing, J.B.: Der Census rauumlicher Complexe: oder Verallgemeinerung des euler’schen Satzes von den Polyedern, vol. 10. Dieterich, Goettingen (1862)

22. Moebius, A.F.: Theorie der elementaren verwandtschaft. Berichte der Saechsischen Akademie der Wissensschaften 15, 18–57 (1863)
23. Blackmore, D., Peters, T.J.: Computational topology, pp. 491–545. Elsevier, Amsterdam (2007)
24. Tourlakis, G., Mylopoulos, J.: Some results in computational topology. Journal of the ACM (JACM) 20(3), 439–455 (1973)
25. Bubenik, P., Kim, P.T.: A statistical approach to persistent homology. Homology, Homotopy and Applications 9(2), 337–362 (2007)
26. Burton, B.A.: Computational topology with Regina: Algorithms, heuristics and implementations, vol. 597, pp. 195–224. American Mathematical Society, Providence (2013)
27. Carlsson, G.: Topology and data. Bulletin of the American Mathematical Society 46(2), 255–308 (2009)
28. Dey, T.K., Edelsbrunner, H., Guha, S.: Computational topology. Contemporary Mathematics 223, 109–144 (1999)
29. Dunfield, N.M., Gukov, S., Rasmussen, J.: The superpolynomial for knot homologies. Experimental Mathematics 15(2), 129–159 (2006)
30. Cerri, A., Fabio, B.D., Ferri, M., Frosini, P., Landi, C.: Betti numbers in multidimensional persistent homology are stable functions. Mathematical Methods in the Applied Sciences 36(12), 1543–1557 (2013)
31. Ghrist, R.: Barcodes: the persistent topology of data. Bulletin of the American Mathematical Society 45(1), 61–75 (2008)
32. Edelsbrunner, H., Morozov, D., Pascucci, V.: Persistence-sensitive simplification functions on 2-manifolds. In: Proceedings of the Twenty-Second Annual Symposium on Computational Geometry, pp. 127–134. ACM (2006)
33. Kaczynski, T., Mischaikow, K., Mrozek, M.: Computational homology, vol. 157. Springer (2004)
34. Pascucci, V., Tricoche, X., Hagen, H., Tierny, J.: Topological Methods in Data Analysis and Visualization: Theory, Algorithms, and Applications (Mathematics+Visualization). Springer, Heidelberg (2011)
35. Robins, V., Abernethy, J., Rooney, N., Bradley, E.: Topology and intelligent data analysis. In: Berthold, M., Lenz, H.-J., Bradley, E., Kruse, R., Borgelt, C. (eds.) IDA 2003. LNCS, vol. 2810, pp. 111–122. Springer, Heidelberg (2003)
36. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science 290(5500), 2319–2323 (2000)
37. Zomorodian, A.: Topology for computing, vol. 16. Cambridge University Press, Cambridge (2005)
38. Holzinger, A., Malle, B., Bloice, M., Wiltgen, M., Ferri, M., Stanganelli, I., Hofmann-Wellenhof, R.: On the generation of point cloud data sets: the first step in the knowledge discovery process. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 57–80. Springer, Heidelberg (2014)
39. Holzinger, A., Stocker, C., Peischl, B., Simonic, K.M.: On using entropy for enhancing handwriting preprocessing. Entropy 14(11), 2324–2350 (2012)
40. Mémoli, F., Sapiro, G.: A theoretical and computational framework for isometry invariant recognition of point cloud data. Foundations of Computational Mathematics 5(3), 313–347 (2005)
41. Canutescu, A.A., Shelenkov, A.A., Dunbrack, R.L.: A graph-theory algorithm for rapid protein side-chain prediction. Protein Science 12(9), 2001–2014 (2003)
42. Salton, G., Wong, A., Yang, C.: A vector space model for automatic indexing. Communications of the ACM 18(11), 620 (1975)

43. Holzinger, A.: Biomedical Informatics: Computational Sciences meets Life Sciences. BoD, Norderstedt (2012)
44. Wagner, H., Dłotko, P., Mrozek, M.: Computational topology in text mining. In: Ferri, M., Frosini, P., Landi, C., Cerri, A., Di Fabio, B. (eds.) CTIC 2012. LNCS, vol. 7309, pp. 68–78. Springer, Heidelberg (2012)
45. Cannon, J.W.: The recognition problem: what is a topological manifold? Bulletin of the American Mathematical Society 84(5), 832–866 (1978)
46. Zomorodian, A.: Chapman & Hall/CRC Applied Algorithms and Data Structures series. In: Computational Topology, pp. 1–31. Chapman and Hall/CRC, Boca Raton (2010), doi:10.1201/9781584888215-c3.
47. Carlsson, G.: Topological pattern recognition for point cloud data (2013)
48. Epstein, C., Carlsson, G., Edelsbrunner, H.: Topological data analysis. Inverse Problems 27(12), 120201 (2011)
49. Aurenhammer, F.: Voronoi diagrams a survey of a fundamental geometric data structure. ACM Computing Surveys (CSUR) 23(3), 345–405 (1991)
50. Epstein, C., Carlsson, G., Edelsbrunner, H.: Topological data analysis. Inverse Problems 27(12) (2011)
51. Zomorodian, A.: Topological Data Analysis, vol. 70, pp. 1–39 (2012)
52. Blumberg, A., Mandell, M.: Quantitative homotopy theory in topological data analysis. Foundations of Computational Mathematics 13(6), 885–911 (2013)
53. Tourlaki, G., Mylopoul, J.: Some results in computational topology. Journal of the ACM (JACM) 20(3), 439–455 (1973)
54. Kong, T.Y., Rosenfeld, A.: Digtial topology - introduction and survey. Computer Vision Graphics and Image Processing 48(3), 357–393 (1989)
55. Holzinger, A., Schantl, J., Schroettner, M., Seifert, C., Verspoor, K.: Biomedical text mining: State-of-the-art, open problems and future challenges. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 271–300. Springer, Berlin (2014)
56. Washio, T., Motoda, H.: State of the art of graph-based data mining. ACM SIGKDD Explorations Newsletter 5(1), 59 (2003)
57. Jiang, C., Coenen, F., Sanderson, R., Zito, M.: Text classification using graph mining-based feature extraction. Knowledge-Based Systems 23(4), 302–308 (2010)
58. Melcuk, I.: Dependency Syntax: Theory and Practice. State University of New York Press (1988)
59. Cook, D.J., Holder, L.B.: Substructure discovery using minimum description length and background knowledge. J. Artif. Int. Res. 1(1), 231–255 (1994)
60. Yoshida, K., Motoda, H., Indurkhya, N.: Graph-based induction as a unified learning framework. Applied Intelligence 4(3), 297–316 (1994)
61. Dehaspe, L., Toivonen, H.: Discovery of frequent DATALOG patterns. Data Mining and Knowledge Discovery 3(1), 7–36 (1999)
62. Fischer, I., Meinl, T.: Graph based molecular data mining – an overview. In: SMC, vol. 5, pp. 4578–4582. IEEE (2004)
63. Morales, L.P., Esteban, A.D., Gervás, P.: Concept-graph based biomedical automatic summarization using ontologies. In: Proceedings of the 3rd Textgraphs Workshop on Graph-Based Algorithms for Natural Language Processing. TextGraphs-3, pp. 53–56. Association for Computational Linguistics, Stroudsburg (2008)
64. Yan, X., Mehan, M.R., Huang, Y., Waterman, M.S., Yu, P.S., Zhou, X.J.: A graph-based approach to systematically reconstruct human transcriptional regulatory modules. Bioinformatics 23(13), i577–i586 (2007)
65. Agirre, E., Soroa, A., Stevenson, M.: Graph-based word sense disambiguation of biomedical documents. Bioinformatics 26(22), 2889–2896 (2010)

66. Liu, H., Hunter, L., Keselj, V., Verspoor, K.: Approximate subgraph matching-based literature mining for biomedical events and relations. PLoS One 8(4) (April 2013)
67. Liu, H., Komandur, R., Verspoor, K.: From graphs to events: A subgraph matching approach for information extraction from biomedical text. In: Proceedings of BioNLP Shared Task 2011 Workshop, pp. 164–172. Association for Computational Linguistics (2011)
68. Nicolau, M., Levine, A.J., Carlsson, G.: Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. Proceedings of the National Academy of Sciences of the United States of America 108(17), 7265–7270 (2011)
69. Carlsson, G.: Topology and Data. Bull. Amer. Math. Soc. 46, 255–308 (2009)
70. Zhu, X.: Persistent homology: An introduction and a new text representation for natural language processing. In: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, pp. 1953–1959. AAAI Press (2013)
71. Zhou, X., Han, H., Chankai, I., Prestrud, A., Brooks, A.: Approaches to text mining for clinical medical records. In: Proceedings of the 2006 ACM Symposium on Applied Computing, SAC 2006, p. 235–239. ACM Press, New York (2006)
72. Corley, C.D., Cook, D.J., Mikler, A.R., Singh, K.P.: Text and structural data mining of influenza mentions in Web and social media. International Journal of Environmental Research and Public Health 7(2), 596–615 (2010)
73. Chen, H., Sharp, B.M.: Content-rich biological network constructed by mining PubMed abstracts. BMC Bioinformatics 5(1), 147 (2004)
74. Barabási, A., Gulbahce, N., Loscalzo, J.: Network medicine: a network-based approach to human disease. Nature Reviews Genetics 12(1), 56–68 (2011)
75. Delfinado, C.J.A., Edelsbrunner, H.: An incremental algorithm for betti numbers of simplicial complexes on the 3-sphere. Computer Aided Geometric Design 12(7), 771–784 (1995)
76. Delfinado, C.J.A., Edelsbrunner, H.: An incremental algorithm for betti numbers of simplicial complexes. In: Proceedings of the Ninth Annual Symposium on Computational Geometry, pp. 232–239. ACM (1993)
77. Ellis, G.: Homological Algebra Programming. Contemporary Mathematics Series, vol. 470, pp. 63–74. Amer Mathematical Soc., Providence (2008)
78. Dumas, J.G., Gautier, T., Giesbrecht, M., Giorgi, P., Hovinen, B., Kaltofen, E., Saunders, B.D., Turner, W.J., Villard, G.: Linbox: A generic library for exact linear algebra. In: Cohen, A.M., Gao, X.S., Takayama, N. (eds.) 1st International Congress of Mathematical Software (ICMS 2002), pp. 40–50. World Scientific (2002)
79. Singh, G., Memoli, F., Carlsson, G.: Topological methods for the analysis of high dimensional data sets and 3d object recognition. In: Botsch, M., Pajarola, R. (eds.) Eurographics Symposium on Point-Based Graphics, vol. 22, pp. 91–100. Euro Graphics (2007)
80. Kobayashi, M.: Resources for studying statistical analysis of biomedical data and R. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 183–195. Springer, Heidelberg (2014)
81. Tausz, A., Vejdemo-Johansson, M., Adams, H.: Javaplex: A research software package for persistent (co) homology (2011), `http://code.google.com/javaplex`
82. Vegter, G.: Computational topology, pp. 517–536. CRC Press, Inc., Boca Raton (2004)
83. Volodin, I., Kuznetsov, V., Fomenko, A.T.: The problem of discriminating algorithmically the standard three-dimensional sphere. Russian Mathematical Surveys 29(5), 71 (1974)

84. Brehm, U., Khnel, W.: Combinatorial manifolds with few vertices. Topology 26(4), 465–473 (1987)
85. Sarkaria, K.S.: Heawood inequalities. Journal of Combinatorial Theory, Series A 46(1), 50–78 (1987)
86. Otasek, D., Pastrello, C., Holzinger, A., Jurisica, I.: Visual Data Mining: Effective Exploration ofthe Biological Universe. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 19–33. Springer, Heidelberg (2014)
87. Holzinger, A.: Human Computer Interaction & Knowledge Discovery (HCI-KDD): What is the benefit of bringing those two fields to work together? In: Cuzzocrea, A., Kittl, C., Simos, D.E., Weippl, E., Xu, L. (eds.) CD-ARES 2013. LNCS, vol. 8127, pp. 319–328. Springer, Heidelberg (2013)
88. Morozov, D., Weber, G.: Distributed merge trees. In: Proceedings of the 18th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, vol. 48, pp. 93–102 (August 2013)
89. Rieck, B., Mara, H., Leitte, H.: Multivariate data analysis using persistence-based filtering and topological signatures. IEEE Transactions on Visualization and Computer Graphics 18(12), 2382–2391 (2012)