ABSTRACT

BENDER, KRISTINA JANE. The Sense of Shape: Discovering Spatial Features of Data with Applied Topology. (Under the direction of Dr. Radmila Sazdanovic.)

This thesis provides an overview of topological data analysis, in particular the process of persistent homology, and focuses on two applications: the first to natural images, and the second to natural language. The mathematical background and the examples reviewed are presented within a framework of conceptual metaphor, specifically mathematical idea analysis, in order to invite readers to consider a humanistic perspective on the components and the potential of applied topology, a dynamic, recently developed discipline.

The Sense of Shape: Discovering Spatial Features of Data with Applied Topology

by

Kristina Jane Bender

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Mathematics

Raleigh, North Carolina

2016

APPROVED BY:

Dr. Ernest Stitzinger

Dr. Michael Grimwood

Dr. Radmila Sazdanovic
Chair of Advisory Committee

BIOGRAPHY

Kristina is a North Carolina native and long-time student whose passion divides evenly between the humanities and mathematics. After graduating from the NC School of Science and Mathematics in 2008, she studied mathematics and literature at UNC Asheville, completing her BA in 2012. She entered the MA in English program at NC State, finishing in 2014, and then returned to mathematics at NC State. She will graduate with an MS in mathematics in May 2016.

ACKNOWLEDGEMENTS

I wish to thank my advisor, Dr. Radmila Sazdanovic, for her guidance on this project and her enthusiastic teaching in two years of topology courses. I also thank Dr. Ernest Stitzinger for his wonderfully challenging algebra classes and all his help throughout the mathematics program. I would also like to thank Dr. Michael Grimwood for his invaluable writing instruction throughout all my years in graduate school. Finally, a special thanks to my family and friends for their love and support.

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

1.1. *Motivation*

Consider the concept of space from a qualitative point of view: space without quantitative or geometric properties, including exact locations, sizes, proportions, and distances. In the context of topology, this consideration affords mathematicians flexibility for understanding the essence of objects in space and the possible transformations among them. Furthermore, this qualitative conception of space, as it captures information relevant to structures throughout most branches of mathematics, functions as a great unifying idea that appears in many domain-specific contexts. Although topology originated as a notoriously pure mathematical enterprise, it has evolved into a discipline rich in practical applications that harness the descriptive power of its combined geometric, combinatoric, and algebraic machinery.

Consider also the means by which observations of various phenomena yield insights about their underlying structure by illuminating decisive features—significant properties that characterize a thing or process. We encounter this operation often, perhaps below the level of conscious awareness when, for example, we infer facts about the world through sensory data. A painting in the style of impressionist pointillism can reveal a full, continuous scene from small, discrete marks of color; we can recognize a person in a crowd by isolating familiar features of the face and body from the visual noise of other people. By mediating visible light, our visual perception enables us to represent and determine the shapes of structures and spaces that surround us.

But some intricate phenomena can elude human perception and thus require that we make intentional decisions concerning our focus of attention. The level of address—the resolution at which we observe and collect information—activates a delicate, sometimes tenuous relationship between two components: first, the specific methodological parameters of data gathering, processing, and analysis, and second, the results of such actions, conclusions which may or may not harbor misleading artifacts, relics of these choices in perspective rather than features inherent in the data.

The generation of high-dimensional, complex, noisy data is a widespread phenomenon, enabled in part by the proliferation of technology and powerful computational tools; furthermore, the attributes of this data—its sheer volume, rate of production, and variety of forms and sources—create formidable obstacles to traditional methods of analysis. In recent years, applications of topology to data analysis in a variety of fields have expanded with notable success. Topological data analysis provides methods for handling the complexity of big data that complement well-established techniques in other domains, including statistics and machine learning, because of three important properties: coordinate invariance, stability under deformation, and compressed representation. A motivating idea is that data has shape, that shape is meaningful, and that topological procedures can uncover qualitative global features of data—its spatial manifestation. This process enables researchers to detect significant properties of the shape underlying the data, like connectedness and holes of various dimensions, and also to visualize it in a way amenable to human understanding and interpretation.

This research encourages a question about meaning: What significant information or insights can topology yield about point cloud data? Many applications of topology to data analysis, including the technique of persistent homology, show successful results that other, more traditional forms of analysis have not. Recognizing that these methods are a relatively recent development in applied topology, practitioners have carefully addressed the meaning and robustness of the interpretive and representational capabilities of these techniques. The current body of research and experimentation involves a number of salient questions: What kind of data appropriately lends itself to applied topological analysis? What advantages does topology offer in certain contexts that other methods do not? How can researchers more adeptly handle the data and interpret the results of topological analysis?

One commonly used technique, persistent homology, begins with the transformation of point-cloud data into nested sequences of simplicial complexes by way of various methods of filtration, which trace, across changes of the proximity parameter, the evolution of a topological structure, as key features appear and disappear during the filtration. Topological features that persist through changes in scale become visible through homology calculations; furthermore, the changes in Betti numbers—the ranks of homology groups—yield useful information about significant features, and they are represented by Betti barcodes for each homology group as a means of visualizing the birth–death intervals of these features across parameter change. Persistent homology thus transforms discrete data sets into combinatorial objects analyzable by methods from algebraic topology, and these objects serve as a visual tool for illuminating the shape of the space.

Any introduction to persistent homology, as one mode of topology applied to data analysis, must address a crucial question: How should mathematicians translate quantitative data (particularly point-cloud data) into the types of spaces that topology handles from a qualitative perspective, using long-established geometric and algebraic processes within the discipline? A collection of points in n-dimensional Euclidean space—an unordered sequence of points, each point being an n-tuple of elements from the space—may represent data of high dimension, in that each point encodes a specific quality within each coordinate of its n dimensions.

For illustration, consider that responses to a survey with multiple (say n) questions may appear as a point cloud, each point expressing the n answers provided by any individual respondent. Some point clouds, such as those collected from analyses of three-dimensional objects, may be approachable and intuitive for humans to ″see″ and ″read,″ but in data of high dimensions, the size and complexity of the raw information often obscures significant features. Persistent homology can transfer such data into a topological context in order to extract coarse global features.

Persistent homology provides a way to characterize the changes that happens within a nested family of simplicial complexes as the proximity parameter increases (that is, it approximates point cloud data with many complexes, rather than one single structure), and Betti barcodes provide a visualization of the lifespans of various features across the changes in scale. In this context, choosing one particular scale of address (i.e., one specific proximity parameter) is unnecessary, as persistent homology offers a way not only to characterize the shape of the data at many values of resolution, but also to record the features that survive for significant intervals of proximity values as well as features that appear only briefly.

In contrast with methods of analysis that attempt to handle noise initially by smoothing or regularizing data, persistent homology facilitates a measure of potential noise without altering the data in which it appears. Furthermore, this method enables analysts to determine the resolution at which a topological feature is relevant, which can provide a nuanced understanding of the data despite its complexity.

1.2. *Results*

The first goal of this study is to explore the mathematics of persistent homology with an overview and discussion of the geometry and algebra enabling its application. The next objective is to highlight two exciting examples of applications to very different disciplines: one involves natural image statistics, and the other involves a novel approach to natural language processing. The final purpose of this work—perhaps its most original contribution to the critical conversation about topological data analysis—is to examine the rhetoric surrounding applied topology through the lens of “mathematical idea analysis,” a perspective pioneered by Lakoff and Nunez in their publication *Where Mathematics Comes From: How the Embodied Mind Brings Mathematics into Being* (*WMCF*), in which they employ cognitive science and conceptual metaphor studies to highlight the influence of embodiment on mathematical thought.

Many well-founded criticisms of the authors' perspective include useful revisions, particularly with regard to how people learn and develop mathematical knowledge and intuition. But no study of their theory exists that contextualizes the potential of these reformulations in the domain of topology, or specifically in the field of topological data analysis. This new perspective on applied topology is appropriate not solely because researchers in this field employ language and informal explanations that connote (and encourage) metaphorical ways to understand its formal mathematical components. It is also fitting because the combination of spatial intuition (geometry) and rigorous calculation (algebra) inherent in applied topology uniquely facilitates a nuanced, precisely contextualized way to address mathematical idea analysis.

This approach to topological data analysis—an approach sensitive to the role metaphor plays in how mathematicians explain and interpret it—sheds light on two intricately connected phenomena. The first is how abstract notions of "shape" derived from embodied experience facilitate and inform robust and, perhaps, unintuitive methods in mathematical data analysis. The second is how these mathematical methods, in turn, enable researchers to see spatial features of data and to derive insights about the meaning of their shape, meaning which might otherwise remain in the shadow of high dimensionality, complexity, and noise, hidden from the capabilities of human spatial intuition.

Thus the concepts involved in applied topology effect a back-and-forth correspondence between grounded spatial experience and abstract formal mathematics. In other words, the metaphorical mappings go both ways rather than, as the authors of *WMCF* would argue, exclusively transform embodied experience unidirectionally into mathematical ideas. Such multidirectional exchanges exemplify the spirit of mathematics and indeed make possible developments in fields including applied topology that rely on the integration of and cross-fertilization between multiple perspectives. A portrait of topological data analysis from this point of view, enhanced by the mathematical background and a couple applications, illustrates the power of metaphor—a version revised by careful critique of the one offered in *WMCF*—in the communication and practice of this recently developed discipline.

1.3. *Outline*

The background section presents an overview of simplicial complexes, homology, filtrations, persistent homology, and Betti barcodes. As a visualization of the mathematics of this technique, the following image summarizes the process of topological data analysis.

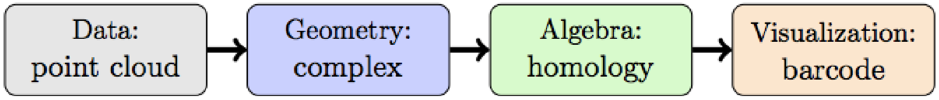


FIGURE 1. The TDA process

Then we explore two very different examples of persistent homology applications: one to natural images and the other to natural language. Finally, a discussion of conceptual metaphor, in the setting of applied topology, revisits the ideas proposed in *WMCF* and restructures its theories to account for the unique metaphorical richness this field exhibits.

2. BACKGROUND

2.1. *A Brief Overview of Simplicial Complexes and Homology*

This section briefly summarizes complexes for the purpose of encouraging geometric intuition, but the reader can explore the resources on which this overview relies for fuller treatments of homology (Carlsson; Ghrist; Hatcher; Nanda and Sazdanovic; Zhu). Simplicial complexes arise from the combination of geometric objects called **simplices**—vertices, edges, triangles, tetrahedra, and their analogs in higher dimensions—according to certain rules of connection.

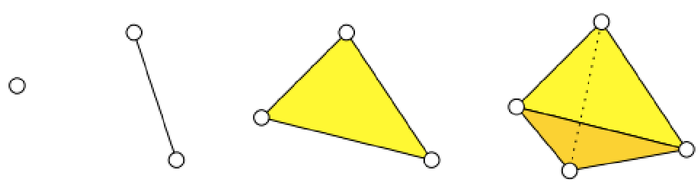


FIGURE 2. Simplices (Zhu)

Notice that the components of simplices are themselves simplices: edges include two vertices and a connecting line, two-dimensional simplices include three vertices, three edges, and a connecting triangle, and so on. The components of a simplex are called **faces**.

Definition 1. *For a set V containing a finite number of vertices v, the* ***simplicial complex*** *K is built from V such that K contains all single-element sets $\{v\}$, all the faces of any simplex in K are also in K, and the intersection of any two simplices σ and τ in K is either $\emptyset$ or a face of both σ and τ. Thus simplices of K join along faces combinatorially.*

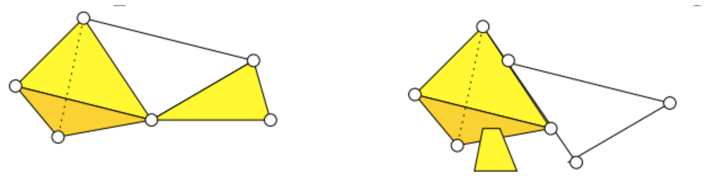


FIGURE 3. Correct and incorrect intersections (Zhu)

Given a complex K, **homology** derives algebraic information about its significant features. Homology computations applied to a single, static simplicial complex yield information about clusters, holes, cavities, etc., and a count of the homology classes in each dimension—the Betti numbers—concisely summarizes those features.

The process of calculating homology groups relies on defining a **boundary map** ∂_d between chain groups $C_d \to C_{d-1}$. The following definitions will clarify these components.

Definition 2. *The i-dimensional **chains** of K are linear combinations of i-dimensional simplices in K (in other words, subsets of these simplices).*

Definition 3. *A **chain group** C_i is the set of i-chains of the simplicial complex K.*

So ∂_d is a linear transformation between these groups, and the kernel and image of this map describe cycles and boundaries, respectively, which are useful for identifying the homology classes. Denote the **cycles** as ker $\partial_d = Z_d$ and the **boundaries** as im $\partial_{d+1} = B_d$. With regard to the edges of a complex, for example, cycles consist of all the possible "rubber bands"—the distinct paths traversing one-dimensional simplices, the edges—and the boundaries consist of the "uninteresting" ones—those that do not surround any holes in the complex. This distinction inspires an intuitive understanding of the following explanation: homology classes come from the removal of trivial paths from the collection of all possible paths in order to reveal significant topological structures.

Definition 4. *The quotient Z_d/B_d gives the d^{th} **homology group**, H_d, for dimension d.*

This process identifies the significant features at each dimension by unveiling the algebraic relation among simplices of every dimension that appear in K. Thus homology separates trivial and nontrivial features, revealing how the simplices of K are connected: for instance, three edges joined on the boundary of a solid triangle yield trivial homology (as the loop can contract over the two-dimensional face), while three edges that form the boundary of a triangle but do not enclose a triangular face will yield nontrivial homology in dimension one (as the loop has no face over which to contract).

Definition 5. *The d^{th} **Betti number** β_d of K is the rank of the homology group H_d.*

The Betti number designates a count of the significant d-dimensional features (again, in dimensions 0, 1, and 2, these are the number of components, loops, and cavities).

Calculating the homology of a given simplicial complex provides powerful insight into the defining features of the space. For complexes built from a set of data points, numerous constructions, such as Vietoris-Rips, Čech, and witness complexes, provide various methods for adding and connecting simplices. These complexes rely on a specified proximity parameter—a choice of distance between vertices in the original set V—that determines whether and how simplices are added to the complex K. But for simplicial complexes that derive from point-cloud data, many nontrivial topological features may not appear at just one parameter value, which generates just one simplicial complex from the data. This limitation highlights the challenge of topologically analyzing raw data in a meaningful way: often no single parameter value captures at once all the significant features underlying the data set.

2.2. *Families of Simplicial Complexes and Filtrations*

Defining persistent homology—which is essentially homology applied to each complex across an entire sequence rather than to a single complex—requires a filtration of the point-cloud space. A filtration corresponds to the application of a chosen distance function that transforms the data into a nested family of complexes. Which complex works best depends on the context and goals of the application: some data sets may be more amenable to certain complex constructions than to others, so researchers must decide which to employ.

Various methods for constructing simplicial complexes employ particular rules for adding simplices to the vertex set. A chosen parameter value ϵ affects the presence of simplices built on a collection of points $\{x_\alpha\}$ in $\mathbb{R}^n$.

Definition 6. *For a **Čech complex**, unordered $(k+1)$-tuples of points $\{x_\alpha\}_{\alpha=0}^{k}$, with closed $\epsilon/2$-ball neighborhoods that have a common point of intersection, determine the k-simplices added to the complex for each ϵ.*

Definition 7. *Alternatively, the **Vietoris-Rips complex** adds k-simplices when unordered $(k+1)$-tuples of points $\{x_\alpha\}_{\alpha=0}^{k}$ are pairwise within the distance parameter ϵ.*

Čech complexes require precise distances among points, resulting in a computational burden. But they are more faithful to the topology: by the *nerve theorem*, a Čech

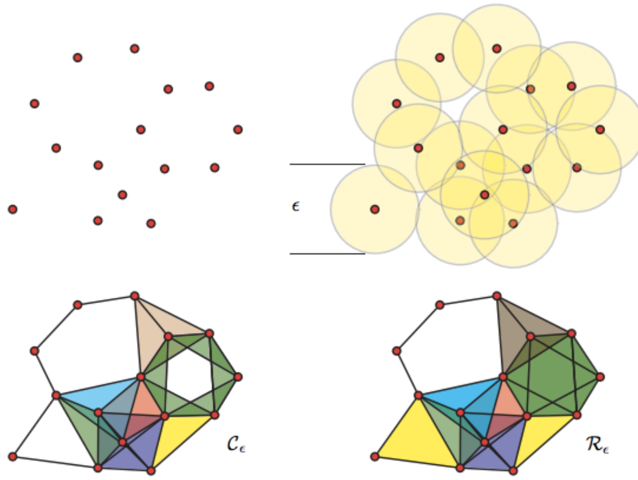


FIGURE 4. Two complexes (Ghrist)

complex is homotopy equivalent to the union of closed balls of radius $\epsilon/2$ about the points of the vertex set, $\{x_\alpha\}$. Thus, despite the potentially high dimension of these complexes, they behave like subsets of the n-dimensional space in which the points originally live. In contrast, while Vietoris-Rips complexes need only coarse proximity information on pairs, their geometric representations might be just an approximation. Both methods often produce simplices of dimensions higher than that of the space. A third case addresses this last problem, but its faithfulness to the space is uncertain.

Definition 8. *A **witness complex** uses a landmark set $\mathcal{L} = \{l_i\} \subset \{x_\alpha\}$ as the vertex set, a subset of points from the space, and adds k-simplices spanned by $\{l_{i_0}, ..., l_{i_k}\}$ if and only if there is a witness point $x \in \{x_\alpha\}$ such that for all j, $d(x, l_{i_j}) \leq m_x + \epsilon$, where $\epsilon > 0$ is the proximity parameter and m_x is the minimum distance from any point in the landmark set to x.*

For a simplicial complex K and a monotonic function f, one can derive a filtration, or sequence of complexes, and a corresponding sequence of homology groups that indicate the changes in topological features of the space across the family of complexes.

Definition 9. *In a **filtration** of K, the function $f : K \to \mathbb{R}$ is non-decreasing along increasing chains of faces, since whenever σ is a face of τ, then $f(\sigma) \leq f(\tau)$. Also, the sublevel set $K(a) = f^{-1}(-\infty, a]$ is a subcomplex of K for all $a \in \mathbb{R}$. If m is the number of simplices in K, then we have $n+1 \leq m+1$ different subcomplexes in an increasing sequence $\emptyset \subseteq K_0 \subseteq K_1 \subseteq ... \subseteq K_n = K$. Furthermore, if $a_0 = -\infty$ and $a_1 < a_2 < ... < a_n$ are function values of simplices in K, then $K_i = K(a_i)$ for all i.*

Since this construction involves adding simplices to the complex in chunks, we find a natural **inclusion map** from the space of K_i to the space of K_j for all $i \leq j$, and consequently an induced **homomorphism between homology groups**, $f_p^{i,j}: H_p(K_i) \to H_p(K_j)$ for each dimension p.

As the sequence of chain complexes arises from increasing the distance parameter ϵ, the homology of the inclusions generates information unavailable from computing the homology of each individual complex alone. Between adjacent subcomplexes K_{i-1} and K_i, homology classes may appear, merge, or die. This is the power of **persistence**—to identify features that survive over a significant parameter range and features that are short lived. While features with small lifespans sometimes indicate noise, this interpretation does not hold universally; for some data, short-lived features may indicate significant structure. The geometric step, defining a function on a topological space, and the algebraic operation, turning the function into measurements, generate information about the resolution or scale of meaningful topological features.

2.3. *Persistent Homology and Betti Barcodes*

Persistence complexes and persistent homology groups derive from the nested sequence of complexes in the chosen filtration, built from and dependent on the point-cloud data set, with an increasing parameter ϵ and the inclusion maps that naturally arise in the construction.

Definition 10. *A **persistence complex** is a sequence of chain complexes $\mathcal{C} = (C_*^i)$ with chain maps $x : C_*^i \to C_*^{i+1}$.*

Definition 11. *Additionally, for $i < j$, the (i,j)-**persistent homology** of $\mathcal{C}$, written $H_*^{i \to j}(\mathcal{C})$, is the image of the induced homomorphism $x_* : H_*(C_*^i) \to H_*(C_*^j)$.*

A classification theorem for persistence modules, which follows from the Structure Theorem for PID's, facilitates a natural interpretation of homology generators. Note that for field $\mathbb{F}$ coefficients, the Structure Theorem for PID's yields the classification of $\mathbb{F}[x]$-modules, since the only graded ideals of $\mathbb{F}[x]$ take the form $x^n \cdot \mathbb{F}[x]$. The following theorem provides an algebraic construction to interconnect the simplicial homology computations of every complex in the filtration (persistent homology groups), thus linking topological features across the filtration and unveiling persistence through algebraic structure (Ghrist).

Theorem. *For a finite persistence module $\mathcal{C}$ with coefficients in a field $\mathbb{F}$,*

$$H_*(\mathcal{C}; \mathbb{F}) \cong \bigoplus_i x^{t_i} \cdot \mathbb{F}[x] \oplus \left(\bigoplus_j x^{r_j} \cdot (\mathbb{F}[x]/(x^{s_j} \cdot \mathbb{F}[x])) \right).$$

In other words, $H_*(\mathcal{C};\mathbb{F})$ indicates the lifespans of homology groups through the exponents of x. The first summand corresponds bijectively to generators born at parameter t_i that persist for all future values, while the second summand corresponds to generators born at parameter r_j that disappear at $r_j + s_j$. This theorem algebraically classifies homology generators of $\mathcal{C}$ with intervals for their birth and death, indicating both transitory features and features that do not disappear as the parameter approaches infinity. These parameter intervals provide a means to visualize the homology generators of $H_p(\mathcal{C};\mathbb{F})$, as the rank of $H_p^{i,j}(\mathcal{C};\mathbb{F})$ is the number of intervals in the barcode of $H_p(\mathcal{C};\mathbb{F})$ that span the parameter range $[i,j]$.

Definition 12. *The p^{th}* ***persistent Betti number*** *is the rank of this group: $\beta_p^{i,j} = rank\ H_p^{i,j}$.*

Definition 13. *A* ***Betti barcode*** *is a graphical representation of $H_p(\mathcal{C},\mathbb{F})$ such that the persistence parameter ϵ increases along the horizontal axis, the vertical axis corresponds to an arbitrary ordering of homology generators at each dimension p, and the horizontal line segments indicate, by their length and starting and ending points, the lifespan and the exact parameter interval of the homology classes to which they correspond.*

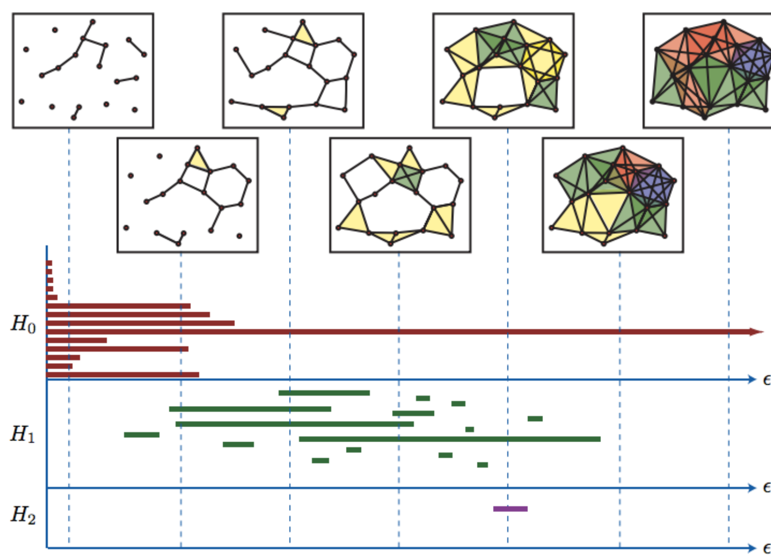


FIGURE 5. A filtration and its barcode (Ghrist)

The barcode qualitatively characterizes ephemeral features with shorter intervals and persistent features with longer intervals, and it does so while simultaneously indicating the parameter range within which those features survive.

3. SOME APPLICATIONS

3.1. *Natural Image Statistics*

One particularly compelling application of persistent homology is to a space of natural images. Carlsson and colleagues follow up on the work by Lee et al. in their study of 3×3 patches of black-and-white images from the database assembled by van Hateran and van der Schaaf, a collection of more than $4,000$ photographs of indoor and outdoor scenes (Carlsson; Carlsson et al.; Ghrist).

FIGURE 6. Some photographs from the database (Perea and Carlsson)

Lee et al. begin by selecting high-contrast, normalized patches, each of which lies in 9 dimensions—one greyscale value for every pixel in the 3×3 grid. They then perform dimensionality reduction on the space in $\mathbb{R}^9$, changing to the Discrete Cosine Transform (DCT) basis to map the patches onto the unit sphere $S^7 \subset \mathbb{R}^8$.

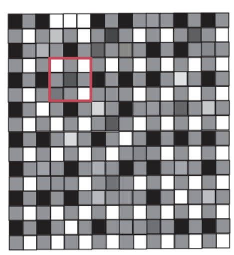


FIGURE 7. A 3×3 pixel sample (Carlsson)

Because the mean intensity normalization and the high-contrast patch selection still result in a space of too large a dimension, Carlsson et al. turn to a focus on density. More specifically, they define the following function.

Definition 14. *The* ***codensity function*** *is obtained in this way: for an integer $k > 0$ and any point x_α in the data, $\delta_k(x_\alpha)$ is the distance in an n-dimensional space from x_α to its k^{th}*

nearest neighbor in the set; thus, δ_k is a positive distribution over the data measuring the radius required to enclose k neighbors in a ball.

These values correlate inversely to the density of the point cloud: larger k-values effect a blurring of finer variations, as the averaging effect among neighbors intensifies. With values k and T, the latter being the upper T-percent of density, they construct a subset parameterized by two values, $\mathcal{M}[k, T]$, a subset that yields a reasonable representation of the significant structure of the data when they choose parameters appropriately.

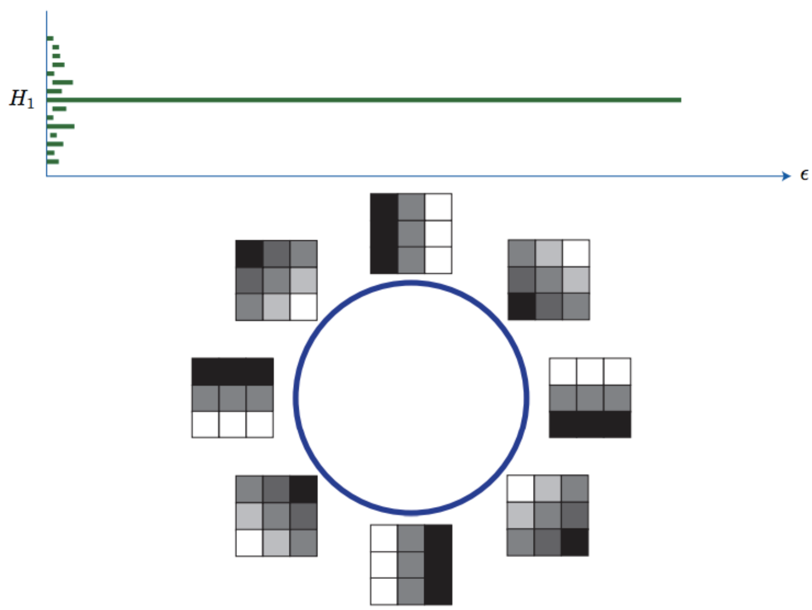


FIGURE 8. The primary circle and barcode (Ghrist)

Using this method of filtration, they derive interesting results for $\mathcal{M}[300, 25]$ and $\mathcal{M}[15, 25]$. After applying these two different density estimations and sampling $5,000$ points from the subsets, they discover that these two dense core subsets have the topology of a circle and a three-circle model, respectively.

What interpretation best explains these homology classes? With $\mathcal{M}[300, 25]$, H_1 exhibits a unique persistence generator in the barcode, so the patches cluster around a primary circle. Upon closer examination, the patterns generating the circle appear to be patches with one dark region, one light region, and a linear greyscale fade from dark to light. As the angle of the transitional line rotates, the most significant patches that emerge exhibit vertical, horizontal, and diagonal black-to-white fading.

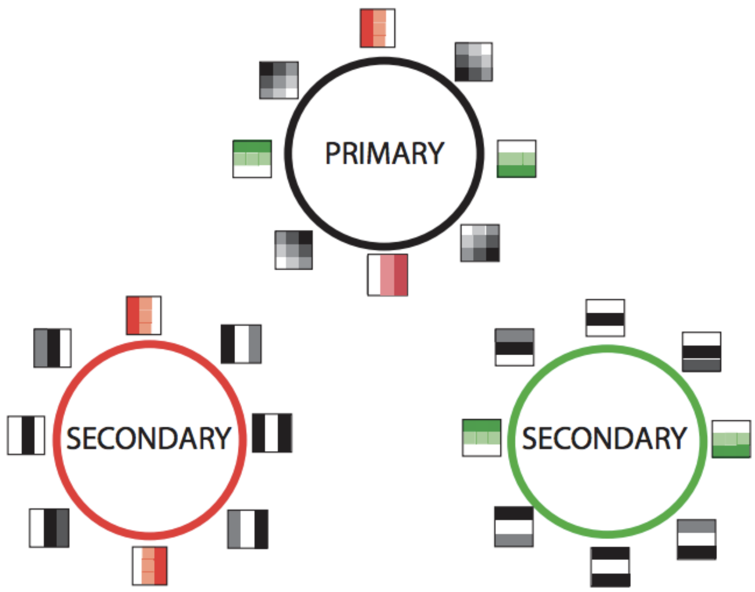


FIGURE 9. The primary and secondary circles (Carlsson)

For the subset $\mathcal{M}[15, 25]$, the persistent H_1 classes yield a Betti number of 5. The k and T values result in finer resolution, so to speak, as less averaging occurs and sensitivity to localized density increases.

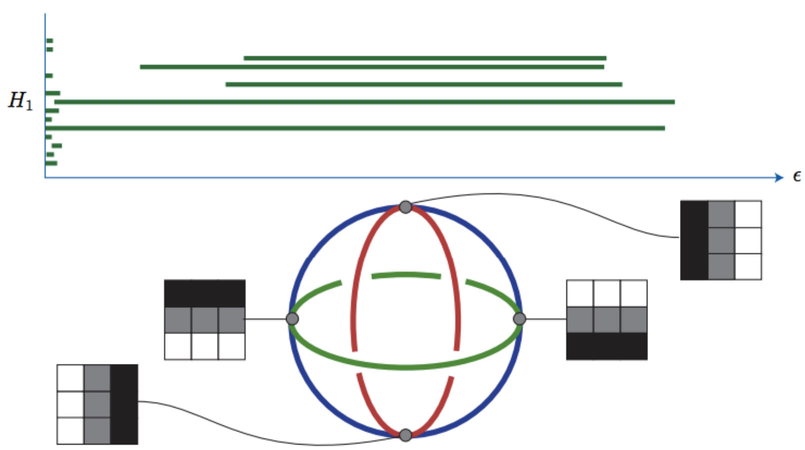


FIGURE 10. The three-circle model and barcode (Ghrist)

Further inquiry supports the presence one connected component, and in addition to the primary circle, two secondary circles appear, each intersecting the primary circle twice but disjoint from each other. The second and third circles correspond to patches with horizontal and vertical orientations, respectively, and along these stratifications

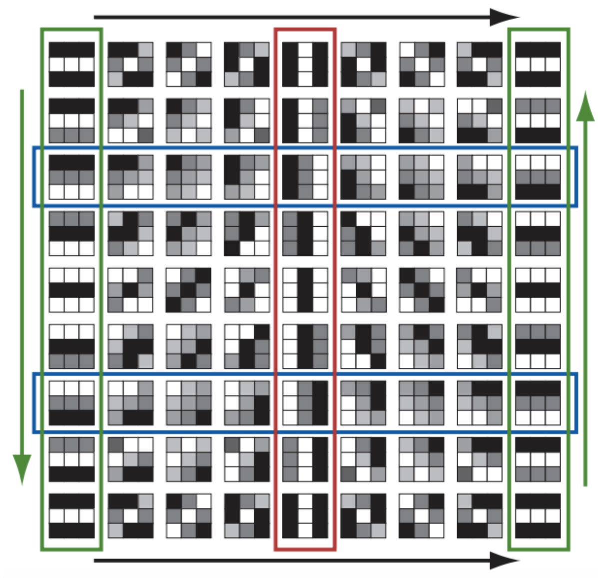


FIGURE 11. The Klein bottle identification diagram (Ghrist)

they encode the transitions between high contrasts of two and three regions within the patches.

Together the three circles form a representation of the shape underlying the relationship between the most densely clustered patches: the Klein bottle. This surprising discovery, illustrated by an indentification diagram, uncovers the transitions between patches clustered around each of the three circles. The polygonal representation captures the relationship between the patches with linear-gradient fades from black to white and the patches with horizontal and vertical orientations of three-patch colorings. Potential applications of this approximation of the space include image compression, among many others.

3.2. *Computational Linguistics and Natural Language Processing*

Although topological data analysis generally applies to the sciences, researchers have recently extended TDA to language. While the study of language has a long history of interdisciplinarity, with statistics and machine learning as two prominent tools for modeling and analyzing language, the proliferation and complexity of data in the form of language suggests that topological data analysis, in combination with current techniques, could provide the machinery to work with this information in novel ways.

However, various theories and practices of text representation and analysis require that researchers make assumptions concerning the investigation of language: What scale of address is appropriate for each approach, and which features of language deserve emphasis? The most obvious and problematic aspect of this research is the question of interpretation: What information does applied topology yield about point-cloud representations of language?

Linguistic analysis might focus on syntax, semantics, discourse structure, authorship, stylistics, text generation, parsing, artificial intelligence development, part-of-speech tagging, network creation, translation, topic and sentiment analysis, or visualization, among other possibilities. Furthermore, the material under consideration might be a single text, a collection of texts from individual authors, genres, or communities, or a massive corpus. In addition, the messy, irregular nature of language—various verb conjugations or noun forms, for example—requires that significant preprocessing regularize the language data to improve the representational or analytical process; tokenization, lemmatization, and stop-word deletion, among other operations, serve this purpose. Researchers could look at word frequency (essentially producing a histogram), break the text into n-grams (strings of n words), graph a word network (based on proximity), or investigate many other phenomena. Thus a key question emerges: How should researchers represent a particular unit of language as data?

As the myriad variables above suggest, a major challenge in this line of research concerns the transformation of text into an appropriate form of quantitative data that, when processed mathematically, yields robust and useful information about the original language source. In particular, in the context of topology, where the results of TDA (applied to any field) often present challenges in interpretation, the path from language-data collection to meaningful interpretation is tenuous.

One application of persistent homology by Zhu breaks texts into segments of a specified number of words, comparing the proximity of strings based on the number of words in common, in an effort to determine the progression of a text and its discourse structure. A major open question is the meaning of homology groups, particularly higher-dimensional ones: For constructed models of language, what exactly do Betti numbers count? And subsequently, what scale of address—local or global—will facilitate the most significant results? We explore this study in depth and analyze its representational and analytical methods.

Applying persistent homology to natural language processing, Zhu provides novel structural representations of text, particularly regarding semantic similarity and "tie-backs" in documents. The two specific tools in this study are the Similarity Filtration (SIF) and Similarity Filtration with Time Skeleton (SIFTS) algorithms, which formalize the concept of curves and loops connecting bag-of-words vectors, in an n-dimensional point cloud, built from divisions of a text's vocabulary. In addition to breaking ground in applications of topology to natural-language data, this work introduces persistent homology as a potentially powerful tool for research in artificial intelligence, specifically feature extraction and representation. While geometric methods, especially those from differential geometry, can facilitate visualizations of information flow in documents, this study offers an algebraic method, applying persistent homology to Vietoris-Rips filtrations of text data.

The intuition behind this application is that documents exhibit a variety of structures expressing information flow, as some might "tell a straight story while others twist and turn." Dividing a document into small units, $x_1, ..., x_n$, and applying a distance function $D(x_i, x_j) \geq 0$, similar units—that is, units sharing many words—will show spatial similarity as vectors in a space, the dimension of which corresponds to the size of text units.

again	all	and	bitsy	came	climbed	down	dried	itsy	out	rain	spider	spout	sun	the	up	washed	water
0	0	0	1	0	1	0	0	1	0	0	1	1	0	2	1	0	1
0	0	1	0	1	0	1	0	0	1	1	1	0	0	2	0	1	0
0	1	1	0	1	0	0	1	0	1	1	0	0	1	2	1	0	0
1	0	1	1	0	1	0	0	1	0	0	1	1	0	2	1	0	0

Figure 12. Vectors for a nursery rhyme (Zhu)

Zhu applies two different algorithms for filtration of the data (SIF and SIFTS), and he interprets the resulting 0^{th} and 1^{th} homology classes, or clusters and holes, respectively. This focus on lower-dimensional persistent Betti numbers typifies topological data analysis, as higher-dimensional homology classes remain difficult to understand and interpret. After describing the two algorithms, Zhu applies them to nursery rhymes, stories, and child and adolescent writing, and then discusses the meaning of the results.

The SIF method ignores the order of text units $x_1, ..., x_n$, so the filtration does not account for the proximity of units within the document, or how the text "flows" from beginning to end. SIF begins with $D_{max} = max\, D(x_i, x_j), \forall i, j = 1, ...n$, or the maximum distance between any two text units in the set. Then for $m = 0, 1, ..., M$, which implements the increase in diameter just like the ϵ parameter explained previously, simplices appear in the Vietoris-Rips complex if two points are within $\frac{m}{M} D_{max}$. The progression of this process allows looser semantic ties between text units across the filtration, which then undergoes persistent homology computations.

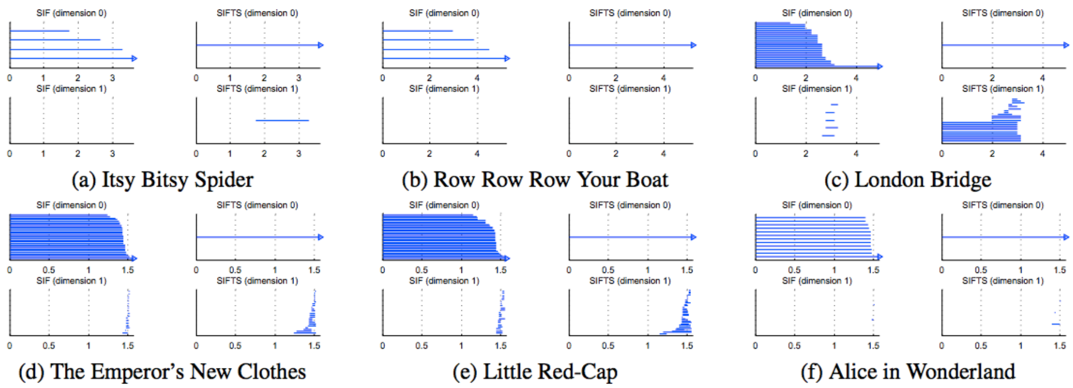
The SIFTS method implements a time component, connecting adjacent text units with "time edges" in the complex, (x_i, x_{i+1}) for $i = 1, ..., n-1$, before the similarity filtration begins. Thus document order determines the initial complex with this preprocessing step: $D(x_i, x_{i+1}) = 0$ for $i = 1, ..., n-1$. Since a time-skeleton edge may be arbitrarily long (if, for example, two adjacent units are dissimilar in their vocabularies), the SIFTS algorithm may uncover semantic tie-backs between the beginning and end of the document under examination (intuitively, this pattern may suggest that a segment of writing revisits its original semantic content, for one reason or another). Note that the time skeleton always makes the SIFTS complex one connected component, since the adjacent units across the document join by time edges, so $\beta_0 = 1$ in every case.

FIGURE 13. Comparison of SIF and SIFTS algorithms (Zhu)

The first illustration involves nursery rhymes—small samples that normally exhibit repetition and are familiar to many people. After minimal tokenization (case-folding, which makes all letters lowercase, and punctuation removal), Zhu separates the rhymes into sentence units and converts them into bag-of-words count vectors. The filtrations proceed over $M = 100$ steps. We characterize the results here.

- *Itsy Bitsy Spider:* SIFTS detects a hole between sentence 1 and 4, in which the spider crawls up the water spout; SIF does not detect this tie-back.
- *Row, Row, Row Your Boat:* Since the four sentences contain distinct words, no holes appear, and the SIFTS algorithm tracks only the linear progression of the rhyme.
- *London Bridge is Falling Down:* Composed of 48 sentences and multiple exact repetitions (12 instances of "my fair lady") and near-repetitions ("Build it up with a and b," a and b being different materials, such as wood, clay, silver, and gold), the SIFTS filtration detects 11 holes in its time skeleton (immediately from preprocessing), while SIF does not, and both algorithms find the latter semantic similarities, showing holes later in the filtration that indicate the near-repetition.

For longer documents, where the units consist of paragraphs or chapters, the textual preprocessing is more involved (including Penn Treebank tokenization and SMART stopword removal, in addition to the case-folding and punctuation removal performed on the rhymes). These vectors undergo cosine similarity and angular distance calculations: $D(x_i, x_j) = \cos^{-1}\left(\frac{x_i^T x_j}{\|x_i\| \cdot \|x_j\|}\right)$. As before, the SIFTS algorithm identifies more holes in general and locates them sooner than SIF. The recurrence of salient words produces longer homology classes. *Little Red-Cap*, for example, shows connections between similar sentences as with, for example, "The better to see you with, my dear" and "The better to eat you with!"



(a) Itsy Bitsy Spider (b) Row Row Row Your Boat (c) London Bridge

(d) The Emperor's New Clothes (e) Little Red-Cap (f) Alice in Wonderland

FIGURE 14. Barcodes for rhymes and stories (Zhu)

The study concludes with an example from the LUCY corpus of child and adolescent writing: the child-writing set includes 150 essays from writers of ages 9–12, with average article length of 11.6 sentences, and the adolescent-writing set includes 48 essays from undergraduate writers with 25.8 sentences on average per document. Zhu hypothesizes that structural richness will increase with age, that is, that more 1-homology classes will appear in the second group.

Computing the SIFTS results, Zhu finds two statistics for comparative ease: $|H_1|$, the total count of 1^{st} homology classes over the entire ϵ parameter range (i.e., the number of bars that appear in the entire barcode), and ϵ^*, the smallest parameter at which a hole first appears in the filtration (this number is set to $\pi/2$, the largest possible angular distance, if none forms).

Calculations show that 87% of children's texts contains loops, compared to 100% of adolescents' texts, with child writers averaging 3 holes and adolescent writers averaging 17.6; furthermore, holes appear earlier in adolescent documents. To control for document length (the adolescents' documents were, on average, twice as long as the children's documents), he computes these statistics for truncated adolescent writing (taking only the first 11 sentences): 98% of texts in the truncated set contains holes, and they contain one more hole on average, but the first appearance of holes shows no statistically significant difference from children's texts.

	child	adolescent	adol. trunc.
holes?	87%	100%*	98%*
$\lvert H_1\rvert$	3.0 ($\pm$0.2)	17.6 ($\pm$0.9)*	3.9 ($\pm$0.2)*
ϵ^*	1.35 ($\pm$.02)	1.27 ($\pm$.02)*	1.38 ($\pm$.01)

FIGURE 15. Child and adolescent writing results (Zhu)

The author's discussion following these experiments with persistent homology includes the argument that this application may be a useful addition to current methods of discourse modeling, parsing, and document comparison in NLP, as homology reveals significant structural features and meaningful differences between text documents. Although Zhu argues that persistent homology with SIFTS does not just count repeated units, future applications and nuanced interpretation could improve understanding of the mechanisms and advantages of topological analyses of natural-language text data.

4. **DISCUSSION**

The applications of persistent homology to natural images and natural language demonstrate the successes of applied topology in handling large point-cloud data sets. They also exemplify the interchange between spatial intuition and mathematical abstraction at the heart of this discipline. Although we are able to encounter images and language directly, analyzing large volumes of data in these forms requires more rigorous ways of looking, more finely tuned scales of resolution. While the geometric and algebraic operations performed on the point cloud data—through the mechanisms of applied topology—may take us into more abstract territory, these transformations yield information about significant spatial features that we can visualize and interpret, enabling us to intuit and explain meaningful structure in the data.

In a survey of biological applications of topology, Nanda and Sazdanovic address the use of both geometry and algebra in TDA: "In order to precisely understand the objects which encode and catalog such [topological] features, we must turn to algebra. Any reader who experiences moral qualms about our descent from the Olympus of geometric shapes to the Hades of algebraic formalism stands in distinguished company" (13). They then provide a clever, compelling quotation of Sir Michael Atiyah: "Algebra is the offer made by the devil to the mathematician. The devil says: 'I will give you this powerful machine, it will answer any question you like. All you need to do is give me your soul: give up geometry and you will have this marvelous machine.'"

Although this pact-at-the-crossroads interpretation of the geometry-to-algebra transition humorously animates a struggle many mathematicians undergo, it also refers to a nontrivial compromise that Lakoff and Nunez discuss at length in *Where Mathematics Comes From*: the exchange of the intuitive and familiar for the abstract and foreign. The goal of this discussion is to introduce and critique the theories presented in *WMCF*, in the hope that revising their ideas about metaphor in mathematics may usefully create a more nuanced perspective on the lively, multifaceted conceptualizations encountered in topological data analysis.

Lakoff and Nunez build their main thesis with ideas from cognitive science and metaphor studies, namely that human embodiment determines the mental processes through which conceptual systems emerge and evolve. In other words, the human body, including the brain, mediates and filters conscious experience in the world, and thus necessarily structures reason within a framework of physical grounding. The authors contend that although people may not always be fully aware of the inextricable connection between embodiment and cognition, this link nonetheless characterizes what, why, and how we know, from birth onward, and specifically explains the ways in which concrete, physical experiences scaffold—or even make possible—abstract understanding.

The authors of *WMCF* use conceptual metaphor as the explanatory mechanism for this pervasive human negotiation between the concrete and the abstract: fundamental aspects of the experiences of human beings, such as sensation, perception, and movement, constitute a basis for the more complex, ethereal elements of human existence, such as language, emotion, knowledge, and culture. They say mathematics, like every other domain of human thought, develops through conceptual metaphor, a universal phenomenon by which humans harness essential, neurophysiological experience to work as a source domain that enables the abstract objects and processes of mathematics to develop and be shared in the cultural imagination.

The program of *WMCF* includes explicit mappings between physical source domains and their elementary targets (“grounding metaphors”) and between these preliminary targets and even more abstract, sophisticated mathematical ideas (“linking metaphors”). For example, through a metaphor built from geometry, meaning may be understood as location in space, and similarity as spatial proximity. The authors then attempt to apply these conceptual metaphors to a wide range of mathematical ideas to demonstrate that their approach is sufficiently powerful to explain the entire discipline as a metaphorical construct contingent upon human embodiment.

The critical responses provide constructive revisions of *WMCF*, reformulations which aim not only to amend its errors and assumptions, but also to align its analytical energies with the lived experiences of students and researchers in mathematics. In his criticism of *WMCF*, Auslander assesses the theory—specifically the idea that all mathematical ideas should be understood as grounded in metaphors from embodied experience—with overall skepticism and a dash of sarcasm: “Mathematical concepts, once they are developed, acquire a life of their own and are dealt with directly. It’s difficult for me to conceive of a metaphor for a real number raised to a complex power [i.e., Euler's identity], but if there is one, I'd sure like to see it" (367).

Schiralli and Sinclair—in their experience-focused critique—want the theory to take into account "the experiences of mathematicians and mathematical learners" (80). Their point is that the authors' "reduction of abstract concepts to more concrete ones through metaphor fails to explain the fundamental processes involved in acts of abstraction" (82). Schiralli and Sinclair suggest that two factors—the variety of meaning-making strategies in mathematics, and metaphors that develop "to and from the ongoing experiences (including non-mathematical ones) of the mathematician"—call into question the thesis of *WMCF* (85). Thus embodied mathematics can instead be interpreted as a continuously evolving "cognitive infrastructure" that facilitates the processes of knowledge creation, such that metaphor serves as a critical tool in the progressive, cumulative, and recursive processes of mathematical learning. A few major themes emerge in the scholarly response to Lakoff and Nunez: the importance of considering cognitive development and ideation in learners, the inability of conceptual metaphor to capture the complexity of mathematical understanding, and the distortion *WMCF* imposes on the real experiences and perspectives of participants in the community.

The chapter entitled "Calculus Without Space or Motion: Weierstrass's Metaphorical Masterpiece" appears just before the conclusion of *WMCF*, which describes the theory and philosophy of embodied mathematics. This chapter features the culmination of Lakoff and Nunez's version of the story of the "discretization program that shaped modern mathematics," an extended treatment of how metaphor "has permitted a reconceptualization of the continuous in terms of the discrete" (260). The authors explain that the "geometric paradigm" was the prevailing understanding of calculus as dealing explicitly with continuous motion through space and change over time—a conceptual framework they deem more natural and intuitive. The movement away from this paradigm was perfected by Weierstrass's program, at the core of which was "a concept that involved no motion (just logical conditions), no continuous space (just discrete entities), no points (just numbers), with functions that are not curves in a plane (but, rather, sets of ordered pairs of numbers)" (309).

Furthermore, Weierstrass's "clever use of conceptual metaphor" meant that he "did not eliminate geometry at all...he just hid it" from view, and in any case, according to Lakoff and Nunez, the geometry metaphorically implicit in classical mathematics could not be left behind anyway (309, 313, 315). The authors argue that although Pierpont "was both attracted and repulsed by the program—attracted by the prospect of increased rigor, repulsed by the loss of tools of visualization and spatial intuition" (as "he knew better" than to trust blindly in formalism), the authors mourn that "he can't help himself" and ultimately "has been converted" (307, 321-22). But Lakoff and Nunez seem to appropriate Pierpont's notions of intuition and arithmetical clarity to serve their goals, rather than fairly recognizing that what Pierpont found "appalling"—the "total separation from the world of our senses" that arithmetization might entail—does not suggest that he believed mathematical insight and curiosity should or would be supplanted by rigor, or that formalism could finally deliver mathematicians to a literal, essentialist truth supplanting embodied human cognition (322).

Pierpont acknowledges that invoking "metaphysical speculations" in mathematics happens more easily in geometric contexts because humans are intimately familiar with embodied experiences of space and motion and are predisposed to view mathematical concepts through the lens of those experiences. But his understanding of mathematics as a "transparent science," particularly as it is transformed by arithmetic rather than geometric formulations, does not preclude "intuition" as Lakoff and Nunez define it—as the ability of humans to rally ideational faculties to facilitate and enhance understanding (322). Certainly humans develop a kind of "natural" familiarity with change and motion within time and space; water flows, balls bounce, plants grow. But Pierpont above all highlights the tension—a difficulty many mathematicians experience—between the clarity and precision of mathematics as one kind of knowledge (which necessarily involves many kinds of intuition, as human minds encounter mathematics), and the tendency of humans (as embodied creatures) to attempt to understand mathematical ideas through physically familiar notions, which can conflict with (and even prevent) accurate understandings of mathematical knowledge.

Whatever one's beliefs about essences in mathematics, rigor within mathematical systems—that is, relational consistencies and the connections between patterns they admit—does not preclude imaginative engagement with those systems or, for that matter, creative translations between private ideations within those systems. Participating successfully in the mathematics community necessarily involves engagement in symbolic, linguistic, and pictorial modes of conceptualization, among many others, and as Pierpont seems to suggest, this participation requires that one abandons neither intuition nor formal standards.

Mathematicians may wonder how Lakoff and Nunez would have analyzed topology, had they included it as a topic in their book. Topology, the study of space under continuous deformations, developed primarily from geometric topics (for example, Euler's seven-bridges problem and the study of knots), and it has evolved over the past century into an important branch in abstract algebra. Topology is notable for its unification of key concepts that appear in different forms in other subdisciplines of mathematics; both its formal methodology and its concern with invariant properties of abstract spaces allow it to generalize ideas that appear as more specific instantiations in other fields.

The authors of *WMCF* might have framed the abstract qualities of topology as further alienation from the "natural" senses, especially considering not only its heavy reliance on set theory but also its evolution from geometric general topology to algebraic topology. Thomas addresses the metaphorical source domains of topological concepts, "grounding metaphors like boundary, connectedness, dimension—ideas which, because of their groundedness, could be used informally for a long time before being mathematized" (192). This insight highlights the tension (or connection) between drawing on embodied experience and formalizing mathematical ideas axiomatically, but examining topology according to the critical revisions of the *WMCF* paradigm reveals that rigor in mathematics does not eradicate this connection but rather transforms and expands it.

Recent developments in algebraic topology and topological data analysis illustrate these problems with the methodology of *WMCF*. The algebra of homology groups is undeniably more "abstract" than direct visual treatments of specific shapes: homology associates sequences of algebraic groups and connecting maps to topological spaces as a way of describing the properties of a space in each dimension. For example, homology distinguishes the boundary of a circle from the boundary of a sphere by capturing their respective one-dimensional and two-dimensional "holes." In this way, homology is a rigorous, formalized method for determining topological invariants: it provides another way of seeing important properties beyond the capacity of direct human experience.

Furthermore, applications of topology have provided fascinating results about the physical world humans inhabit. Persistent homology allows researchers to derive meaningful information from high-dimensional data (such as point clouds from natural-image data and natural-language corpora), as the important "spatial" features, which the scale and noise of the data might obscure in statistical analysis, are revealed through a more holistic, big-picture search for topological invariants. Of course, mathematical applications require sensitive, cautious interpretation, but mathematicians might argue that the evolution of formal mathematics actually enhances the embodied experiences, conceptual potentials, and metaphorical capacities of humans. Applying algebraic topology, specifically persistent homology, to data sets can illuminate significant features of those spaces and open pathways for mathematicians—guided by spatial intuition derived from embodied experience—toward deeper understanding, insight, and discovery.

Students and researchers engage in individual ideational processes as they learn and inquire, and topological applications exemplify the variety of mathematical conceptualizations available in every discipline, as well as the opportunities for translation between diverse representations of objects and processes. Thus the application of conceptual metaphor, according to Lakoff and Nunez's program, might not reveal the multidirectionality of conceptual abstraction (not captured by strict "physical source" to "abstract target" delineations for metaphorical mappings) and the multiplicity of mental constructions—influenced as much by embodiment as by mathematical experience—that participants variously invoke and transform in their negotiations with shared conceptual systems. *WMCF* certainly explores relevant, insightful perspectives on mathematical ideas, but in analyzing those ideas as generated by static, universal metaphors, the authors neglect the changeable, unique cognitive processes of individuals and the ways formal mathematics invites (rather than precludes) multifaceted translations of concepts.

Returning to topological data analysis, we see that the concept of shape motivates abstract mathematical analyses of data in search of spatial features. Although the meaning of these features and of their persistence across algebraic transformations may depart from familiar embodied meanings (shape as pixel configurations in images or as semantic similarity in language), the axiomatized mathematical approach of persistent homology facilitates and even scaffolds our interpretations of these meanings: metaphorical conceptualizations of the shape of data, grounded in the formal mathematics of topology and abstract algebra, enable us to qualitatively encounter spaces once hidden in the high dimension and complexity of quantitative data.

REFERENCES

[1] Auslander, J. Embodied mathematics. American Scientist 89(4): 366–367 (2001).

[2] Carlsson, G. Topology and data. Bulletin of the American Mathematical Society 46(2): 255–308 (2009).

[3] Carlsson, G., Ishkhanov, T., De Silva, V., and Zomorodian, A. On the local behavior of spaces of natural images. International Journal of Computer Vision 76(1): 1–12 (2008).

[4] Edelsbrunner, H., and Harer, J. Computational topology: an introduction. American Mathematical Society (2010).

[5] Ghrist, R. Barcodes: the persistent topology of data. Bulletin of the American Mathematical Society 45(1): 61–75 (2008).

[6] Hatcher, A. Algebraic topology. Cambridge University Press, Cambridge (2002).

[7] Lakoff, G., and Nunez, R. E. Where mathematics comes from: how the embodied mind brings mathematics into being. Basic Books, New York (2000).

[8] Lee, A. B., Pedersen, K. S., and Mumford, D.B. The nonlinear statistics of high-contrast patches in natural images. International Journal of Computer Vision 54(1–3): 83–103 (2003).

[9] Nanda, V., and Sazdanovic, R. Simplicial models and topological inference in biological systems. Discrete and Topological Models in Molecular Biology, Natural Computing Series, Springer: 109–141 (2014).

[10] Nilsson, D., and Ekgren, A. Mapper and betti 0 barcodes applied to random indexing word-spaces: a first survey. BS thesis, Stockholm (2013).

[11] Perea, J.A., and Carlsson, G. A klein-bottle-based dictionary for texture representation. International Journal of Computer Vision 107(1): 75–97 (2014).

[12] Sahlgren, M. The word-space model: using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. PhD thesis, Stockholm (2006).

[13] Schiralli, M., and Sinclair, N. A constructive response to where mathematics comes from. Educational Studies in Mathematics 52(1): 79–91 (2003).

[14] Thomas, R. Idea analysis of algebraic groups: a critical comment on George Lakoff and Rafael Nunez's *Where Mathematics Comes From*. Philosophical Psychology 15(2): 185–195 (2002).

[15] Zhu, X. Persistent homology: an introduction and a new text representation for natural language processing. Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, AAAI Press: 1953–1959 (2013).