Learning From the Shape of Data

Sarita Rosenstock

1 Introduction

Data scientists take large quantities of noisy measurements and transform them into tractable, qualitative descriptions of the phenomena being measured. While this frequently involves statistical methods, the burgeoning field of data science distinguishes itself from statistics by branching out to a wider range of methods from mathematics and computer science. One such distinctly non-statistical method of growing popularity is topological data analysis (TDA). Topology is the study of the properties of shapes that are invariant under continuous deformations, such as stretching, twisting, bending, or re-scaling, but not tearing or gluing. TDA aims to identify the essential "structure" of a data set as it "appears" in an abstract space of measurement outcomes. This paper is an attempt to reconstruct the reasoning given by data scientists as to why and how the resulting analysis should be understood as reflecting significant features of the systems that generated the data.

In section 2 I describe TDA in detail. Section 3 discusses what I take to be the central feature of the success of TDA, and section 4 examines the role of spatial reasoning in TDA, and how it can give us insight into the connections between the data scientists' and philosophers' notions of "structure."

2 Topological data analysis

The phrase "topological data analysis" is used to refer to a variety of data science practices that use tools from algebraic topology to make inferences about the "shape" of data clouds as they appear in the "space" of possible observations. For now, the term *data* refers to a set of real vectors corresponding to a series of observations. This is an adequate definition for capturing natural language use of the term, but one might object that it does not necessarily capture what data *is*. One of the goals of TDA is to circumvent some of the arbitrariness involved in presenting data as real vectors. A *data cloud* can thus be thought of as a visual representation of this set of vectors as "points" in a (high dimensional generalization of) space. But in what space? The abstract "space" where data lives is generally some form of *metric space*, or set X of points (including at least the data points) together with a notion of "distance" $d(\ , \ )$ between the points. For example, I may have data about the weights of each of a large number of potatoes. The distance between these data points would just be the pairwise difference in weight between two potatoes according to a fixed unit, such as pounds.

A characteristic problem of analyzing large data sets is deciding how to combine many different types of measurements into a shared metric space. I can also add information about the length, color, number of eyes, etc. for each potato, creating an n-dimensional space, where n is the number of potato attributes. The "distance" between two data points is now some combination of the distances given by weights, lengths, color, etc. But how should the notions of distance given by each variable combine into "distance" in the total space of possible variable values? The "standard" way of aggregating one-dimensional metrics into a shared metric space is to imagine each metric as an axis in an n-dimensional Cartesian grid, with distance given by the Cartesian distance as follows. Let $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$ be two sets of potato measurements. Then $d(x, y) = \sqrt{(x_1 - y_1)^2 + ... + (x_n - y_n)^2}$. Setting aside the fact that there are other viable options for constructing distances from these values, notice that this expression does not include units. Should weight be presented in pounds or tons? Of course we know how to translate between these two units, and we consider the choice more of notational convenience than theoretically meaningful. But if we are looking to the "shape" of data for information about the system being measured, the data cloud will look much more "flat" if we use tons rather than pounds. It is thus desirable to consider properties of the data cloud that do not depend on the particular choice of metric space or unit, but which are shared by a variety of plausible modeling choices.

Such considerations motivate the use of *topological*, as opposed to geometric methods. Topology is the mathematical field that studies properties of shapes that remain constant under stretching, twisting, or otherwise deforming. Topologists attend to more general features of metric spaces that would be present under different modeling assumptions, called *topological invariants.* Since data sets are finite, although they may suggest some underlying shape, they likely will not do so uniquely. This is the standard curve-fitting problem in higher dimensions: for any discrete set of points, there are an infinite number of continuous curves (or shapes) that contain (or approximate) the locations of those points. As with the curve-fitting problem, external considerations guide the choice of continuous object, rather than just the bare, uninterpreted set of data points. One may have a priori reasons to expect that the "right" curve is quadratic, for example, or that the modeling goal should be to minimize mean-squared error.

2.1 Clusters

The simplest example of TDA, and the one most broadly used by data scientists generally, is *cluster analysis.* The idea behind cluster analysis is to ask: do my data points naturally divide into sub-categories of data points more similar to one another than the overall space? Such a situation indicates that there is some non-trivial structure underlying the data associated with such groupings, which one may interpret as "natural kinds" in the space. Cluster analysis is in this way closely related to regression analysis—clusters point towards a correlation among variables, one of the main "signals" data scientists hope to read off of large data sets.

Sometimes, external considerations about the type of data under consideration can influence how one chooses to carve a data set into clusters. Even in the absence of such guidance, natural clusters may be easily "seen" when the data is graphed. With larger and higher dimensional data sets to analyze, these heuristics are less useful, and data scientists would prefer a principled algorithmic approach to clustering. This would amount to a function that takes metric spaces (X, d)—here understood as data sets $X = \{x_1, ..., x_n\}$ with a notion of "distance" $d(x_i, x_j)$—as inputs, and outputs *partitions* of that data into clusters of data points that are "close together."

2.2 Constructing Shapes

The most common method to construct a shape from a data cloud is roughly as follows. Enclose each data point in a "ball" of radius ε centered on that point. As ε gets larger, the cloud will cease to look like isolated points and start to gain shape. Once it gets too large, though, we are left with a single shapeless blob. We use this idea to construct a *simplicial complex*, beginning with the data points as vertices.[1] Where 2 balls intersect, we add an *edge* between them. When 3 balls intersect, we add a *face* enclosed by the three edges. This process continues, creating higher dimensional *n-faces* where $n + 1$ balls intersect. The result is called a *Čech complex*.[2]

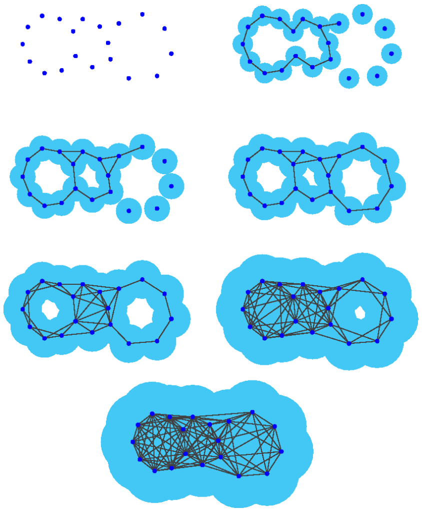

Figure 1: Constructing a Čech complex as ε increases, from Bubenik (2015).

This is an intuitively plausible way to construct a discrete shape from a data cloud. A clustering can be read off of a Čech complex by grouping data points according to whether they are connected in a single component of the complex. This may be complicated by the presence of noise—a single anomalous data point might connect otherwise robustly distinct clusters. This can be side-stepped by either looking at only regions that are highly connected, or avoided altogether by filtering and "cleaning" the data prior to analysis.

[1] See Hatcher (2002) section 2.1 for a precise definition of a simplicial complex.

[2] In practice, TDA employs a more computationally tractable approximation thereof, called a *witness complex.* See Carlsson (2009) section 2 for details.

2.3 Holes and voids

Identifying the clusters of a simplicial complex appears is a special case of a more general phenomenon of *homology*. Homology is a method of classifying shapes by looking at how many "holes" the shape has. No matter how much you stretch and twist it, a circle will always have a "hole" in it, a sphere will always have a *void* or *cavity*, an innertube will always have the "donut hole" as well as a void in the interior that inflates.

In looking at the connected components of a Čech complex, we are considering the H_0-*homology* of the complex (considered as a topological space). We can similarly attend to the H_1-*homology* of the complex by looking for "holes," or the H_2-*homology* by looking at "cells," and so on to higher dimensions with less intuitive interpretations.

Example 1 (Cosmology)**.** van de Weygaert et al. (2011) study the homology of density level sets of an ensemble of randomly generated cosmic mass distributions. They analyze the evolution of H_1, H_2, and H_3-homology over time in n-body simulations, revealing characteristic patterns of different dark energy models. They show how homology can track cosmological structures of independent interest to physicists, such as matter power spectra and non-Gaussianity in the primordial density field.

2.4 Persistence

The motivating idea behind the construction of a Čech complex is that we can imagine data as being uniformly sampled (with noise) from some underlying "shape" in the metric state space, and we can use these data points to infer the global structure of the "object" we are sampling from. The more samples we look at, the more accurate our picture of the shape will be. For sufficiently small ε-balls, the complex will not have any more structure than the bare data set. Similarly, when the balls get too large, there is nothing more to look at than a giant blob. The "right" choice of ε is at some intermediate size, but how should it be chosen? If we chose an ε that is too small, we will get a shape with a lot more holes, disconnected components, etc., than we think are meaningful. In other words, we retain some of the noisy features of the data cloud that we were trying to eliminate. But we risk going to far, and making ε large enough to obscure both noise *and* meaningful information from the data.

A natural way to solve this problem is to look at many different choices of ε, and use external considerations to decide which gives the best resolution of the data shape. Two more problems arise when we do this, though. For one, the whole point of data analysis is to simplify and compress information about a system, and having a variety of different models we can choose from does not simplify matters. Second, there may be different features that arise at different resolutions that are equally significant, and this multi-level picture can get lost if we have to choose a single model among the many possibilities. For example, data may be dense in some regions but sparse in others, where relevant shapes require larger ε-balls to be "seen".

The key insight that unlocked the power of TDA was the idea of "topological persistence," introduced to data analysis in (Edelsbrunner et al., 2002). Briefly: instead of picking a particular resolution to look at, we look at them all, but take advantage of a trick from algebraic topology to connect complexes at different scales in a sophisticated and efficient way The result is the association of a data cloud with a *persistence module* that encodes how the cloud changes structurally as ε increases. Homology is then computed for these modules, and the result is typically expressed as a *homological barcode*, as in figure 2. The "bars" begin when a feature is "born" and end when it "dies." Short intervals in barcodes are often attributed to either measurement noise or inadequate sampling, whereas long, "persistent" bars are thought to reveal real geometric features of the space being sampled from.

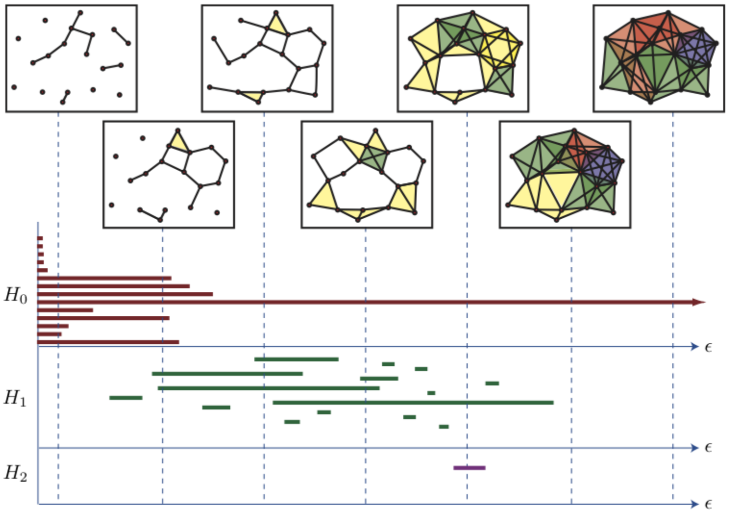

Figure 2: Example of a homological barcode, from Ghrist (2008).

This construction is enabled by a structure theorem of Crawley-Boevey (2015), demonstrating that persistent modules can be uniquely represented as a direct sum of *interval modules*. Not only is this decomposition more computationally tractable to analyze than (sets of) complexes, but the barcode itself provides a visual summary of behavior as ε increases. When the number of features is large, data analysts will also sometime use *persistence diagrams* instead of barcodes. These diagrams plot features on a birth-death axis. See figure 3 for a diagram of voids—H_2-homological features—in a cosmological model from example 1. Dots on the diagonal indicate voids that die quickly after birth, and those farther away are more persistent.

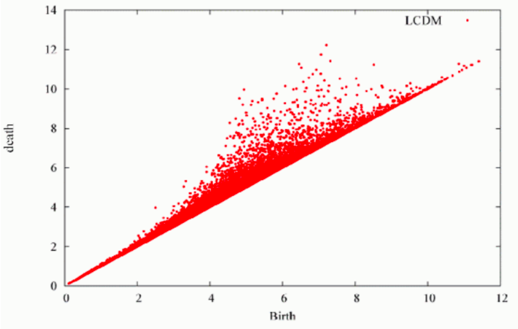

Figure 3: Birth-death diagram of voids in a cosmological model (van de Weygaert et al., 2011).

2.5 Stability

One way to interpret ε is as a modeling parameter, corresponding to the resolution or scale we use to construct a shape from the data cloud. The persistent features of a Čech complex are those that are *stable*, or robust under perturbations of the parameter value. Longer bars in barcodes represent features that appear for a wider range of ε values, indicating that these features are robust and unlikely to constitute mere noise. Cohen-Steiner et al. (2007) made this precise by proving that for a large class of constructions (including Čech complexes), persistence diagrams are *stable*, meaning that small perturbations of the initial data set result in correspondingly small changes in the resulting persistence diagram.

We can use this same method to consider stability across other indexing parameters as well at fixed resolution, as in the following example.

Example 2 (Arteries)**.** Bendich et al. (2016) employ topological data analysis to study the structure of arteries in the human brain. They uniformly sample a large number of points from a blood vessel diagram (weighted by thickness of vessel), and construct a Čech complex from this data cloud, analyzing the H_0 and H_1 persistence diagrams over the growing size of ε-balls in the Čech complex. They look at persistent H_0 over a stack of “horizontal slices” of the artery diagram.

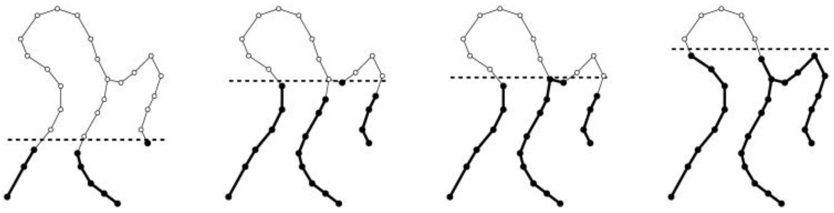

Figure 4: Horizontal slices of the artery diagram, from Bendich et al. (2016).

The authors found significant correlation between certain features of these homological barcodes and the age and sex of the subjects, with the age correlation a significant improvement over previous attempts at analyzing similar data. For example, older brains tended to have the longest bars in the latter barcodes.

We can thus understand persistence modules as assembling a sequence of $(n-1)$-dimensional models in a sequence indexed by an n^{th} parameter, such as resolution or time. Dimensionality reduction is a common feature of data analysis techniques. Data often comes in the form of large vectors, and the goal is often to *compress* them—express as much of the original information as possible with in as few dimensions as possible. This amounts to selecting features or parameters of interest and suppressing the rest in order to highlight general patterns. Reducing data models to 2-3 dimensions also makes them more visualizable, making them more useful to researchers to observe patterns, as well as easier to communicate to the public. Persistence modules provide the benefits of low dimensional visualizability without throwing away the information in the extra dimensions.

Summary

The general procedure for determining persistent homology is as follows.

1. Generate a sequence of shapes (CW-complexes) from the data cloud.
2. Transform the sequence into a *persistence module* indexed by a parameter such as resolution or time.
3. Construct a visual summary of the persistence module as a barcode or diagram.

1 and 3 are straightforwardly motivated—1 from the intuitive geometric interpretation of data as (noisily) sampled from some underlying shape, and 3 from the Crawley-Boevey structure theorem.

3 Functoriality in TDA

Whichever method we use to give shape to our data cloud, the result is a topological space. More specifically, it is a (finitely generated) *CW-complex*: a particularly "well-behaved" topological space that is constructed by "gluing" n-disks along their boundary $(n-1)$-spheres. Čech complexes are CW-complexes, as are all of the other constructions of figures from data clouds that we will consider here.

Homology is a general way of associating, to each of these shapes X built from a data cloud, a (finitely generated) Abelian *homology group* $H_n(X)$. For each group, $H_n(X)$ essentially characterizes how many "holes" are present in each dimension. $H_0(X)$ tracks the connected components, $H_1(X)$ tracks holes, $H_2(X)$ tracks cells, or the number of valves that would be required "inflate" the hollows of the shape. This extends to higher dimensions, but most TDA applications only look at these three, as these are the most spatially intuitive.

In order for persistence analysis to work, we need to be able to track shapes as they appear and disappear when ε increases. Homology is not merely an assignment of a group to each complex that provides information about its shape. Homology is *functorial* in the sense that it comes equipped with a notion of how to translate maps between complexes into maps between groups while preserving all relevant topological information.[3] The functoriality of homology enables us to do three important things, which are essential

[3] This functoriality is inherited from the homology functor from the category of CW-complexes **CW** to the category of abelian groups **Ab** that lies at the heart of algebraic topology.

to its utility in analyzing data: identify local structures, connect complexes as parameters vary, and compare complexes constructed from different samples.

The homology group $H_n(C)$ of a complex C tells us how many "holes" it has, but it does not tell us *where* the holes are, or how big they are. This is to be expected—recall that while these complexes "live" in metric spaces, TDA looks at more general, topological rather than geometric features of them, which are preserved when the space is stretched or rotated. Nonetheless, topological spaces still have a (albeit weaker) notion of "nearness" associated with them. We can cover our topological space with "neighborhoods," and ask, relative to a particular cover, whether a "hole" is contained in a single neighborhood.

So, if there is a feature of interest, we can locate it in a neighborhood $U \subseteq C$ and think of this neighborhood as its own complex. We can then look at the inclusion map $\iota : U \to X$ that just acts as the identity on that neighborhood. Since homology is *functorial*, this induces a corresponding map $\iota_* : H_n(U) \to H_n(C)$, allowing us to track the n-dimensional "hole" in the group $H_n(C)$ as the image $\iota_*(H_n(U)) \subseteq H_n(C)$. (See Zomorodian, Afra and Carlsson, Gunnar (2008) for details on this localization method).

We can thus refer to a particular hole as it appears in the homology group, rather than referring to it spatially. But even more importantly, given a map $f : C \to D$ that identifies two complexes via their underlying metric space, we can ask whether the hole contained in U *persists* under the transformation f by seeing whether $f_*(\iota_*(H_n(U)))$ vanishes. This is what enables the use of homological barcodes to encode information about when holes form and disappear as a complex is constructed in stages by increasing ε. Each bar corresponds to a different hole, understood locally in this way.

Most practitioners will admit that the interpretation of homology in data is unclear. While increasing in popularity of late, TDA is still relatively niche. It is often reserved for situations in which traditional data analysis tools have failed to bear fruit, and TDA is one of many attempts to gain insight into the data—its more of a trial and error situation.

Since persistent homology has these nice properties, data scientists will often shoe-horn questions about data into the shape of a homology problem in order to make it tractable. For example, they might add extra edges to a Čech complex to turn open chains into closed loops. Or they might chose a particular dimensional reduction in which loops arise, as in Perea and Harer (2015). A fun example is the study of "tendrils", another geometric property of data clouds that is of potential interest. See image below—n tendrils emanating from a central cluster. By supplementing TDA with a procedure for identifying such clusters, these can be removed, and the tendrils can be tracked via the persistent H_0-homology of the resulting data cloud. Nicolau et al. (2011) use this technique to classify breast cancer types.

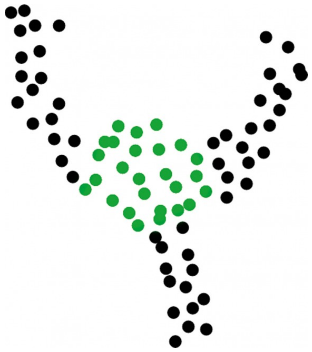

Figure 5: Visualization of data that features "tendrils", from Lesnick (2013).

Data scientists study persistent homology, not because they think of "counting holes" as the right way to characterize data, but rather because it is has really desirable features summarized by its functoriality. While the recent proliferation of these methods might be dismissed as mere hammer-nailing, it should rather be said that since we have very few tools to work with, we had better hope this problem can become nail-shaped.

4 TDA and spatial inference

4.1 Geometric understanding

Though there is disagreement about the nature and pervasiveness of its influence, visual, spatial, and aesthetic intuitions unarguably play a role in science. Rather than sidelining it, topological data analysts explicitly embrace the role of visual intuitions in their scientific work. Again, TDA is a second-line resource for data that is particularly intractable to analyze, which puts creativity at the center of its application.

The goal of data analysis is to identify patterns in data that provide concise, comprehensible summaries of the system that point towards features of significance in broad classes of systems. Such recognition of patterns of sufficient generality without overfitting is the holy grail of artificial intelligence and machine learning research. In the mean-time, scientists still rely heavily on the *je ne sais quoi* features acquired through visual intuition to guide inquiry. To aid the evocations of these intuitions, data scientists will play around with parameters and data filtering. Since spatial intuitions exist at lower dimensions, the ability to use persistence modules to reduce dimensionality without losing information makes it especially useful.

While subjective visual judgments clearly dominate the earlier stages of inquiry, data analysts still return to more traditional empirical methods for post hoc justification. Even if a topological feature is robust under TDA analysis, the real measure of a successful analysis is whether it corresponds to a feature of the system of independent interest to scientists. Patterns found through random applications of TDA might lead scientists to look for such an independently interesting feature of a system, but if one cannot be found, the shapes identified in the data remain merely curiosities. In example 2, if barcodes did not track gender and age but some other feature that we do not independently classify as a natural kind, researchers would likely not have identified it. Even if they had stumbled upon a barcode pattern by chance, it would not have mattered if they could not tell a compelling story about what characteristic the pattern characterizes.

So spatial intuitions play a central role in the context of discovery, while their influence is fortified by the introduction of external empirical considerations at the stage of justification. But they reappear when the results are communicated to others, in the visual summary provided by a homological barcode or diagram. This allows data scientists to again invoke visual intuitions in evaluating the results of the analysis, which now contain all of the information about the persistence of shapes in an easily consumable, two-dimensional aid. The functoriality of TDA carries the visual information in CW-complexes through various reformulations until it finally reappears again in yet another visual format in its presentation.

4.2 Diagrammatic reasoning

Returning from our detour into the cognitive realm, we might wonder how all of this can be incorporated into a formal epistemic story about the structure of topological data models. Here, we can learn much from the vast literature on diagrammatic reasoning in Euclidean geometry. Critics of the rigor of reasoning from diagrams in geometric ‘proofs’ point to the fact that such proofs use a particular illustration to make an inference about all possible illustrations. However, philosophers of mathematical practice have recently come to appreciate the role of diagrams in generating and communicating geometric knowledge. Manders (2008) argues that ancient geometers were careful to rely on diagrams only for demonstrations about what he calls *co-exact* features—those that are relatively insensitive to the range of variation in possible visual representations, such as part-whole and boundary-interior relationships (and of course, homology). Mumma (2010) takes this a step further and develops a formal account of Euclidean proofs that includes both sentential and diagrammatic components.

How does this bear on TDA? Earlier, I noted that data analysts are concerned with ensuring that inferences about data rely only on real structural features of observations, rather than incidental features of how data is embedded in a metric space. At issue is the level of generality one can adopt when making inferences from a single visual representation of data, picked somewhat arbitrarily from an ensemble of possible alternative, equally valid representations. TDA resolves this issue by requiring that the analyzed features of data models be *functorial* with respect to maps that preserve what they take to be the relevant structural features of models, and *persistent* across parameters when the "right" value is not known.

4.3 Structure

The forgoing discussion about TDA hints at a new way to understand the relationship between the following two conceptions of "structure" in models of scientific theories:

1. The relevant causal and explanatory features of a system, abstracted from the noise present in any observation of the system; and

2. The content of a description of a physical system, abstracted from the particular language and formalism used to present it.

The relationship between these two notions at first appears relatively superficial. Yes, they are both ways of getting at what is *really there* in a physical system, but they seem to refer to completely different stages of scientific representation. The first comes in at the stage of observation and experimentation, referring to the structure of a particular physical system under observation. The second relates to extracting information from an idealized model (perhaps constructed out of “cleaned” data from the previous stage). The structure here consists of the components of the model that are actually doing the representational work, rather than merely scaffolding this content in language and symbols.

Scientific practice is a holistic process, and measurement cannot and should not be wholly separated from formal representation. Among other connections, observations inform how systems should be represented, and representations indicate directions for further research and measurement. Of course, no one would say these two senses of structure are mere homonyms, as they are clearly relying on similar intuitions about how structure is supposed go beyond particulars of an instance to something more essential. However, they seem to operate in different realms—structure$_1$ in the physical world and structure$_2$ in Platonic heaven—and thus they surely must be orthogonal to one another.

Nonetheless, I think TDA indicates how these two notions of structure are much closer than one might initially suspect. The first clue to this comes upon noticing that more than being intertwined with one another, particular acts of structural refinement by scientists may exist between realms. For example, suppose I collect demographic data that includes the hair color of participants, and include hair color as a feature of my initial abstract representation of this population, recorded as an RGB hex code. I then decide that precise hair color is not a relevant consideration for the theoretical purpose at hand, so I switch to presenting hair color information more coarsely as either light, medium, or dark. I can think of this “throwing away”, “rounding off” or “smoothing out” as an act of cleaning data, obscuring noise at the observation level, and perhaps fundamentally changing the type of data I collect. Alternatively, I can think of it as refining my model—obscuring noise at the representational level. It amounts to the same act, viewed through different lenses. The moral is that the boundary between data and formalism is not completely clear.

5 Discussion

Data scientists sometimes claim that the functoriality of homology is critical to TDA’s utility in revealing and interpreting structural features of data sets. This paper offers an account of how and why this is this case. There are various reasons to suspect that topological features correspond to meaningful signals in a data set. Moreover, topological features are accessible to visual cognition to aid in scientific interpretation. Since homology is *functorial* relative to the category (**CW**) that delineates the relevant structures, it is ensured that the reasons we had for thinking topological features were meaningful are preserved in the translation from data cloud to homological barcode.

Requiring functoriality constrains the tools that are available to us to analyze data, and homology is particularly well understood mathematically. Data scientists thus often try to apply persistent homology even if it is not immediately obvious why topological features of the data should be important. But by identifying that functoriality is operative in enabling robust inferences in TDA, we can use category theoretic tools to express the general features of *any* data analysis that might be epistemically sufficient. Bubenik and Scott (2014) provide the mathematical tools, and this paper supplements them by demonstrating how to construct an inferential narrative to justify their epistemic value. An obvious next step would be to explore new functorial data analysis methods (or functorializing old methods).

References

Bendich, P., J. S. Marron, E. Miller, A. Pieloch, and S. Skwerer (2016). Persistent Homology Analysis of Brain Artery Trees. *The Annals of Applied Statistics 10*(1), 198–218.

Bubenik, P. (2015). Statistical Topological Data Analysis Using Persistence Landscapes. *Journal of Machine Learning Research 16*, 77–102.

Bubenik, P. and J. A. Scott (2014). Categorification of Persistent Homology. *Discrete & Computational Geometry 51*(3), 600–627.

Carlsson, G. (2009). Topology and Data. *Bulletin of the American Mathematical Society 46*(2), 255–308.

Cohen-Steiner, D., H. Edelsbrunner, and J. Harer (2007). Stability of Persistence Diagrams. *Discrete Comput. Geom. 37*(1), 103–120.

Crawley-Boevey, W. (2015). Decomposition of Pointwise Finite-Dimensional Persistence Modules. *J. Algebr. Appl. 14*(05), 1550066.

Edelsbrunner, H., D. Letscher, and A. Zomorodian (2002). Topological Persistence and Simplification. *Discrete & Computational Geometry 28*, 511–533.

Ghrist, R. (2008). Barcodes: The Persistent Topology of Data. *Bulletin of the American Mathematical Society 45*(1), 61–75.

Hatcher, A. (2002). *Algebraic Topology*. Cambridge UP, Cambridge.

Lesnick, M. (2013). Studying the Shape of Data Using Topology. `https://www.ias.edu/ideas/2013/lesnick-topological-data-analysis`. Accessed: 2018-11-19.

Manders, K. (2008). The Euclidean Diagram (1995). In *The Philosophy of Mathematical Practice*. Oxford: Oxford University Press.

Mumma, J. (2010). Proofs, Pictures, and Euclid. *Synthese 175*(2), 255–287.

Nicolau, M., A. J. Levine, and G. Carlsson (2011). 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 108*(17), 7265–7270.

Perea, J. A. and J. Harer (2015). Sliding Windows and Persistence: An Application of Topological Methods to Signal Analysis. *Foundations of Computational Mathematics 15*(3), 799–838.

van de Weygaert, R., G. Vegter, H. Edelsbrunner, B. J. T. Jones, P. Pranav, C. Park, W. A. Hellwing, B. Eldering, N. Kruithof, E. G. P. p. Bos, J. Hidding, J. Feldbrugge, E. ten Have, M. van Engelen, M. Caroli, and M. Teillaud (2011). Alpha, Betti and the Megaparsec Universe: On the Topology of the Cosmic Web. In M. L. Gavrilova, C. J. K. Tan, and M. A. Mostafavi (Eds.), *Transactions on Computational Science XIV: Special Issue on Voronoi Diagrams and Delaunay Triangulation*, pp. 60–101. Berlin, Heidelberg: Springer Berlin Heidelberg.

Zomorodian, Afra and Carlsson, Gunnar (2008). Localized homology. *Computational Geometry 41*(3), 126–148.