1 Introduction

In the last two or three decades, the need for machine learning and artificial intelligence has grown dramatically. As the tasks we undertake become ever more ambitious, both in terms of size and complexity, it is imperative that the available methods keep pace with these demands. A critical component of any such method is the ability to model very large and complex data sets. There is a large suite of powerful modeling methods based on linear algebra and cluster analysis that can often provide solutions for the problems that arise. Although they are often successful, these methods suffer from some weaknesses. In the case of algebraic methods, it is the fact that they are not always flexible enough to model complex data, such as data sets of financial transactions or of surveys. Clustering methods by definition cannot model continuous phenomena. Additionally, they often require choosing thresholds for which there are no good theoretical justifications. What we will discuss in this volume is a modeling methodology called topological data analysis, or TDA, in which data is instead modeled by geometric objects, namely graphs and their higher-dimensional versions, simplicial complexes. Topological data analysis has been under development during the last 20 or so years, and has been applied in many diverse situations. Its starting point is a set equipped with a metric, typically given as a dissimilarity measure on the data points, which can be regarded as endowing the data with a shape. This shape is very informative, in that it describes the overall organization of the data set and therefore enables interrogation of various kinds to take place; TDA provides methods for measuring the shape, in a suitable sense. This is useful in as much as it allows one to access information about the overall organization. In addition, TDA can be used to study data sets of complex description, which might be thought of as unstructured data, where the data points themselves are sets equipped with a dissimilarity measure. For example, one might consider data sets of molecules, where each data point consists of a set of atoms and a set of bonds between those atoms, and use the bonds to construct a metric on the set of atoms. This idea leads to powerful methods for the vectorization of complex unstructured data. The methodology uses and is inspired by the methods of topology, the mathematical study of shape, and we now give a more detailed description of how it works.

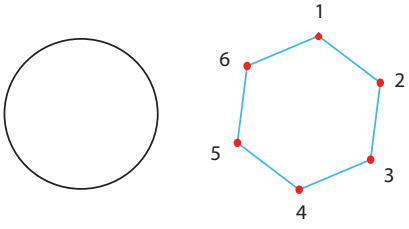
Much of mathematics can be characterized as the construction of methods for organizing infinite sets into understandable representations. Euclidean spaces are organized using the notions of vector spaces and affine spaces, which allows one to arrange the (infinite) underlying sets into understandable objects which can readily be manipulated and which can be used to construct new objects from old in systematic ways.

https://doi.org/10.1017/9781108975704.003 Published online by Cambridge University Press

Similarly, the notion of an algebraic variety allows one to work effectively with the zero sets of sets of polynomials in many variables. The notion of shape is similarly encoded by the notion of a *metric space*, a set equipped with a distance function satisfying three simple axioms. This abstract notion permits one to study not only ordinary notions of shape in two and three dimensions but also higher-dimensional analogues, as well as objects like the p-adic integers, which may not be immediately recognized as being geometric in character. Thus, the notion of a metric serves as a useful organizing principle for mathematical objects. The approach that we will describe demonstrates that the notion of metric spaces acts as an organizing principle for finite but large data sets as well.

Topology is one of the branches of mathematics which studies properties of shapes. The aspect of the study of shapes which is particular to topology can be described in terms of three points.

1. The properties of a shape studied by topology are independent of any particular coordinate representation of the shape in question, and instead depend only on the pairwise distances between the points making up the shape.
2. The topological properties of shapes are *deformation invariant*, i.e. they do not change if the shape is stretched or compressed. They would of course change if non-continuous transformations are applied that "tear" the space.
3. Topology constructs compressed representations of shapes, which retain many interesting and useful qualitative features while ignoring some fine detail.

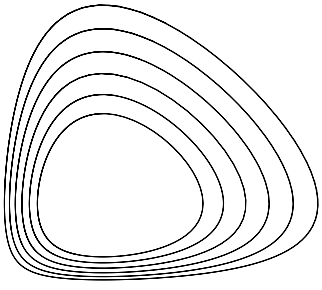


Thus, topology deals with shapes in two distinct ways. The first is by building compressed combinatorial representations of shapes, via processes such as triangulation. Of course some information about a shape is lost in this discretization, such as fine-scale curvature information, but, as in the example above, the rough overall structure is preserved in passing from the circle to the hexagon. The second way is by attempting to measure shapes, or aspects of shapes. This is done via *homological signatures*, which permit a kind of count of the occurrences of patterns within a shape. The adaptation of these signatures to the study of point cloud data (sets of data points in space) is the subject of this book.

A motivating example comes from contemplation of the phase space pictures of the Lotka–Volterra equations. Recall that these equations describe the population dynamics in a simple predator–prey model and result in oscillatory behavior, which gives rise to loops in phase space. One could of course describe such a loop by giving a precise parametrization (i.e. a system of local coordinates) for the loop. However, for many purposes the fact that the shape of the phase portrait is a smooth deformation of a circle is the most salient detail. More generally, consider the family of examples coming from

https://doi.org/10.1017/9781108975704.003 Published online by Cambridge University Press

the mapping of a circle to Euclidean space under a wide variety of embeddings. A salient qualitative description would extract the fact that the underlying data comprised a circle. The intuitive idea behind algebraic topology is that one should try to distinguish or perhaps even characterize spaces by the occurrences of such qualitative patterns within a space. For the Lotka–Volterra example one could say that a characteristic pattern is the presence of a loop in the space surrounding the empty region in the middle. One could say intuitively that the count of loops in the phase portrait is one, in that there is "essentially" only one loop in the space. The same characterization would hold for an annulus, where the essential loop winds around the central removed disc. It is not so easy to make mathematical sense of this observation, because there are often families of loops that we would regard as being essentially the same, as in the figure below.



The presence of essentially one loop is something which a priori is difficult to quantify, since in fact there is an uncountable infinity of actual loops which have the same behavior, i.e. they each wind around the hole once. In order to resolve this difficulty and formalize the notion that there is essentially only one loop, we are forced to perform some abstract constructions involving equivalence relations to obtain a sensible way of counting the number of loops. The idea is that we must regard many different loops as equivalent, in order to get a count of the occurrences, not of each individual loop but, rather, of a whole class of equivalent loops. This step is responsible for much of the abstraction which has been introduced into the subject. Once that layer of abstraction has been built, it provides a way to detect the presence of geometric patterns of certain types. The general idea of a pattern is of course somewhat diffuse, with many different meanings in many different contexts. In the geometric context, we define patterns as maps from a template space, such as a circle, into the space. A large part of the subject concerns the process of reducing the abstract constructions described above to much more concrete mathematical constructions, involving row and column operations on matrices. The goals of the present volume are the following.

- To introduce the pattern detection signatures which come up in algebraic topology, and to simultaneously develop the matrix methods which make them into computable and usable invariants for various geometric problems, particularly in the domain of *point clouds* or *finite metric spaces*. We hope that the introduction of the relevant matrix algorithms will begin to bridge the gap between topology as practiced "by hand" and the computational world. We will describe the standard methods of

https://doi.org/10.1017/9781108975704.003 Published online by Cambridge University Press

homology, which attach a list of non-negative integers (called Betti numbers) to any topological space, and we also discuss the adaptation of homology to a tool for the study of point clouds. This adaptation is called *persistent homology*.

- To introduce the mathematics surrounding the collection of *persistence barcodes* or *persistence diagrams*, which are the values taken by persistent homology constructions. Unlike Betti numbers, which are integer valued, persistent homology takes its values in multisets of intervals on the real line. As such, persistence barcodes have a mix of continuous and discrete structure. The study of these spaces from various points of view, so as to be able to make them maximally useful in various problem domains, is one of the most important research directions within applied topology.
- To describe various examples of applications of topological methods to various problem domains.

1.1 Overview

The purpose of this book is to develop topological techniques for the study of the qualitative properties of geometric objects, particularly those objects which arise in real-world situations such as sets of experimental data, scanned images of various geometric objects, and arrays of points arising in engineering applications. The mathematical formalism called *algebraic topology*, and more specifically *homology theory*, turns out to be a useful tool in making precise various informal, intuitive, geometric notions such as holes, tunnels, voids, connected components, and cycles. This precision has been quite useful in mathematics proper, in situations where we are given geometric objects in closed form and where calculations are carried out by hand. In recent years, there has been a movement toward improving the formalism so that it becomes capable of dealing with geometric objects from real-world situations. This has meant that the formalism must be able to deal with geometric objects given via incomplete information (i.e. as a finite but large sample, perhaps with noise, from a geometric object) and that automatic techniques for computing the homology are needed. We refer to this extension of standard topological techniques as *computational topology*, and it is the subject of this volume.

We will assume that the reader is familiar with basic algebra, groups, and vector spaces.

In this introductory chapter, we will sketch all the main ideas of computational topology, without going into technical detail. The remaining chapters will then include a precise technical development of the ideas as well as some applications of the theory to actual situations.

1.2 Examples of Qualitative Properties in Applications

1.2.1 Diabetes Data and Clustering

Diabetes is a metabolic disorder which is characterized by elevated blood glucose levels. Its symptoms include excessive thirst and frequent urination. In order to understand the

https://doi.org/10.1017/9781108975704.003 Published online by Cambridge University Press

disease more precisely, it is important to understand the possible configurations of values that various metabolic variables can exhibit. The kind of understanding we hope for is geometric in nature. An analysis of this type was carried out in the 1970s by Reaven & Miller (1979).

In this study, a collection of five parameters (four metabolic quantities and the relative weight) were measured for each patient. Each patient then corresponds to a single data point in five-dimensional space. In Reaven & Miller (1979), the *projection pursuit* method was used to produce a three-dimensional projection of the data set, which looks like the situation on the left in Figure 1.1.

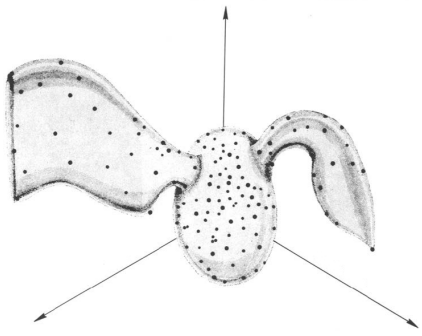


Figure 1.1 Diabetes patient distribution. See the main text for a description of the figure. From Reaven & Miller (1979), reproduced with permission of Springer–Nature, © 1979.

Each patient was classified as being normal, chemical diabetic, or overt diabetic. This is a classification which physicians devised using their observation of the patients. It was observed that the normal patients occupied the central rounded object, and the chemical diabetics and overt diabetics corresponded to the two "ears" in the picture. Another very interesting method for visualizing the set was introduced in Diaconis & Friedman (1980). These visualizations suggest that the two forms of diabetes are actually fundamentally different ailments. In fact, physicians have understood that diabetes occurs in two forms, "Type I" and "Type II". Type I patients often have juvenile onset, and the disease may be independent of the patient's life style choices. Type II diabetes more often occurs later in life and appears to depend on life style choices. The chemical diabetics are likely to be thought of as Type II diabetics who eventually may arrive at the overt diabetic stage, while the overt diabetics might arrive at overt status directly.

In this case, the qualitative property of the figure that is relevant is that it has the two distinct ears, coming out of a central core. Although human beings can recognize this fact in this projection, it is important to formalize mathematically what this means, so that one can hope to automate the recognition of this qualitative property. For example, there may be data sets for which no two- or three-dimensional projection gives a full picture of the nature of the set. In this case, the mathematical version of this statement would be as follows. We suppose that the three categories of patients (normal, chemical, and overt) correspond to three different regions A, B, and C in five-dimensional Euclidean space.

https://doi.org/10.1017/9781108975704.003 Published online by Cambridge University Press

What this experiment suggests is that if we consider the union $X = A \cup B \cup C$ (which corresponds to all patients) and then remove A, the region corresponding to the normal patients, the region we are left with breaks up into two distinct connected pieces, which do not overlap and in fact are substantially removed from each other. Clustering techniques from statistics were used in Symons (1981) to find methods to differentiate between these two components. The qualitative question above about the nature of the disease can now be stated as asking how many connected components are present in the space of all patients having some form of diabetes. Finding the number of connected components of a geometric object is a topological question.

1.2.2 Periodic Motion

Imagine that we are tracking a moving object in space, and that the information is given in terms of a three-dimensional coordinate system, so that we are given coordinates $(x(t), y(t), z(t))$. If we want to know whether the object is moving periodically, say, because it is orbiting around a planet, we can simply check whether the values of the coordinates repeat after some fixed period of time. Suppose, however, that we are not given the time values corresponding to the points but just a set of positions, and want to determine whether the object is undergoing periodic motion. We would thus like to know whether the set of positions forms a closed loop in space. If the object is orbiting around a single planet or star, and we therefore know by Kepler's laws that the geometric shape of the orbit must be that of an ellipse, we can determine that the object is orbiting by simply curve-fitting an ellipse to the data set of positions. Suppose, however, that the object is being acted on gravitationally by several other objects, so that the path is not a familiar kind of closed curve. We would then still want to know whether the space of positions is a closed loop, but perhaps not one for which we have a familiar set of coordinatizations. The qualitative property in which we are interested is whether the space is a closed loop, and we would like to develop techniques which allow us to determine this without necessarily asking for a particular coordinatization of the curve. In other words, we are asking whether our space is a closed loop of some kind, not exactly what type of loop it is.

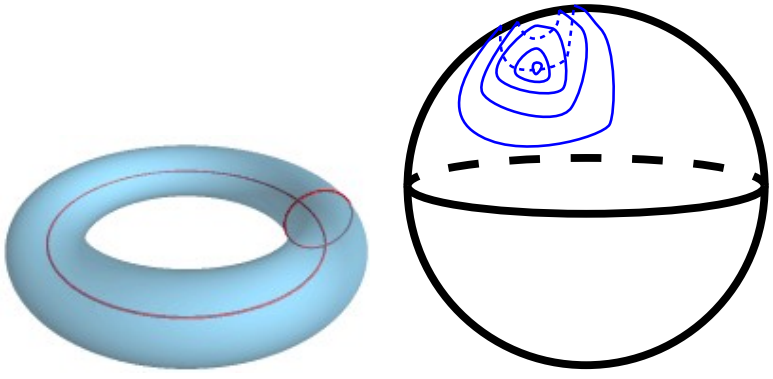
A more difficult situation is where we are not actually given the values of the position of the object but, rather, a family of images taken from a digital camera. In this case, the set of these images actually lies in a very high-dimensional space, namely the space of all p-vectors, where p is the number of pixels. Each pixel of each image is given a value, the gray-scale intensity at that pixel, and so each image corresponds to a vector, with a coordinate for each pixel. If we take many images sequentially, we will obtain a family of points in the p-dimensional space, which lies along a subset which should be identified topologically with the set of positions of the object, i.e. a circle. So, although this set is not identified with a circle through any simple set of equations in p variables, the qualitative information that it is a circle is contained in this data. This is an example of an exotic coordinatization of a space (namely the circle) and shows that, in order to analyze this kind of data, it would be very useful to have tools which can tell whether a space is a closed loop, without its having to be any particular loop. In other words, coordinate-free tools are very useful.

https://doi.org/10.1017/9781108975704.003 Published online by Cambridge University Press

1.2.3 Curve and Shape Recognition

There are situations where we have geometric objects which do not come from experimental data, but where qualitative and coordinate free tools are of value for their analysis. Consider the problem of recognizing hand-printed characters. Hand-printed versions of a particular letter or number can vary a great deal. In fact, there exists a database (the MNIST database Bottou et al. 1994) which comprises many different handwritten versions of the numerals from 0 to 9. The variability comes from the fact that different people develop slightly different versions of the same character, and in fact these versions are sufficiently different that they may sometimes be used to identify the person who wrote them. Differences may also arise from the fact that one may not be looking directly, i.e. head-on, at the paper where the character is drawn, or that it may not be drawn on a flat surface. However, there are a sufficient number of qualitative cues which allow human beings to identify characters despite this variability. For example, if we compare the letter "A" with the letter "B", it is not hard to see that the letter "A" has a single closed loop in it, while "B" has two. Thus, the number of loops is a sufficient criterion to distinguish between these two letters, and it suggests the potential value of developing rigorous and automatic methods for determining the number of loops. Suppose instead that we consider the problem of distinguishing between the letter "U" and the letter "V". In this case, neither letter has a loop, but "V" has a "corner" and "U" does not. This is another useful qualitative cue. Finally, if we attempt to distinguish "C" from "I", we see that neither letter has a loop, and further that there are no corners, but that "C" has a curved arc while "I" does not. This is again a useful qualitative cue, which it will be useful to formalize.

Similar cues can allow us to distinguish between two-dimensional objects in $\mathbb{R}^3$, i.e. to perform *shape recognition*. For example, to distinguish the sphere from the torus (the two-dimensional surface of a doughnut), we can observe that every loop on the sphere can be contracted down to a point, while the torus has two obvious types of loop which cannot.



Similarly, one can distinguish between a tetrahedron and a cube by noting that a cube has eight vertices and 12 edges, while the tetrahedron has four vertices and six edges. Note that both these criteria are robust in the sense that if we make smooth deformations of the objects in question, these characteristics still remain unchanged.

https://doi.org/10.1017/9781108975704.003 Published online by Cambridge University Press

We will see later that cues involving “corners”, “curved arcs”, “vertices”, and “edges” are not directly topological. We will develop methods for recognizing these cues topologically on new spaces that we have constructed from the old ones using tangential information.

https://doi.org/10.1017/9781108975704.003 Published online by Cambridge University Press