TOPOS, and why you should care about it

Robert Adler writes in the first of a new series of columns:

To pure mathematicians, a topos is a type of category that behaves like the category of sheaves on a topological space. (I'll bet that this definition didn't help many IMS members very much. It certainly doesn't 'do it' for me!) To literary buffs, topos describes a traditional theme or motif, or a literary convention. In both cases, the source is the Greek τοποζ, meaning 'place' although in the literary setting the term comes from κοινοζ τοποζ, literally, 'common place'. In Hebrew, the root is ספט, related to climbing or ascension.

However, in this and three more columns to come, *TOPOS* is simply an acronym for *Topology, Probability and Statistics,* and the aim of the columns will be to convince you that by exploiting the theme of TOPOS we are going to be able to ascend to a place where three disciplines are today combining to produce elegant mathematics, powerful statistical tools, and challenges galore.

For some motivational background, let's go back to the 1970s when a topologist by the name of John Tukey (yes! He was trained as a topologist, not as a statistician) introduced and fought for a statistical methodology he called *EDA* (Exploratory Data Analysis). In Tukey's own words, a short description of EDA is that

1. It is an attitude *and*
2. A flexibility *and*
3. Some graph paper (or transparencies, or both).

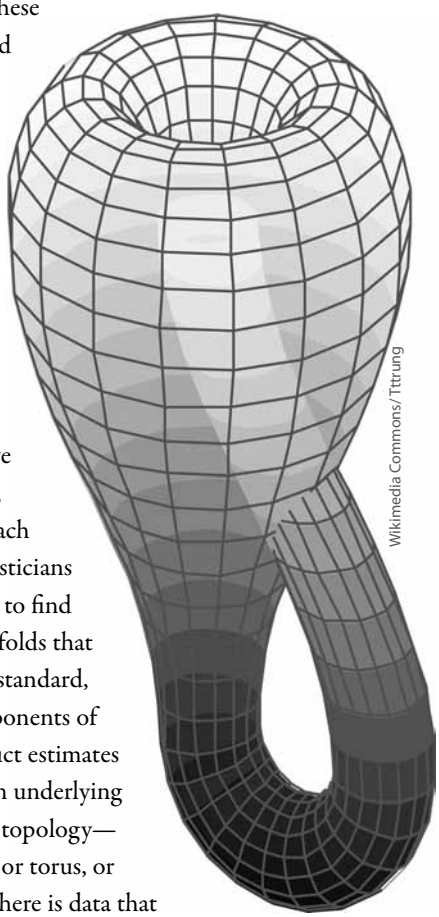
Although it seems hard to believe today, I am old enough to remember that EDA involved a serious challenge to the dominant statistical paradigms of the time, which were based almost solely on hypothesis testing and parameter estimation. The idea that one should play with data first seemed outlandish. Times have changed, and today EDA is not only an established practice, but, backed up by some very nice probability, it enjoys a solid scientific foundation and has provided grist to the mills of theoreticians of many kinds.

Times continue to change, and while Tukey's 'attitude' and 'flexibility' are as relevant today as they ever were, graph paper is hard to find, and transparencies have long since yielded to online presentations. Moreso, even if we had them on our desks, they would be next to useless for EDA-ing the large data sets that are so common today. What is arising as an EDA-like tool, however, is something known as *TDA: Topological* Data Analysis. Perhaps not surprisingly, given the precedent set by Tukey, TDA comes out of the world of topologists rather than the classical analysers of data, statisticians

The appearance of topology—and especially algebraic topology—as a tool for understanding real world problems is actually not surprising. With data arriving in greater and greater numbers and, in particular, in higher and higher dimensions, we need number crunchers for the numbers and understanding for the dimensions. The people who have invested most effort over the last half century or so thinking about the structure of high-dimensional objects are algebraic topologists, and some of the braver ones have been stepping outside their homological ivory towers at the peak of pure mathematics to turn their insight and the powerful tools they have developed to non-mathematical uses.

Indeed, there has been such a significant expansion of activity in applying algebraic topology that the term *Applied Algebraic Topology* is no longer the oxymoron that it would have been a decade ago. Real applications of the techniques of algebraic topology are already appearing (in part, thanks to the enormous success of the programs *Topological Data Analysis* and *Sensor Topology of Minimalist Planning [SToMP]* funded by the US Defense Advanced Research Projects Agency (DARPA) as well as similar, but smaller, European programs).

Many of the applications in TDA are of the dimension reduction and manifold learning type. These problems are far from new, and both statistics and computer science are awash with algorithms for doing this well and efficiently. Many of these, such as projection pursuit and principal component analysis, have lead to deep mathematical problems demanding serious statistical and probabilistic analyses, so IMS members have had a lot to do here. However, TDA adds a very novel approach to these problems. While statisticians and their friends typically like to find estimated subspaces and manifolds that are close to the truth in some standard, quantifiable distance, the proponents of TDA look for ways to construct estimates *that have the right topology*. An underlying theme is to get the qualitative topology—does the data live on a sphere, or torus, or maybe Klein bottle (and yes, there is data that lives on a Klein bottle, but that is for a later column)—right *before* starting quantitative analysis. Indeed, this is very



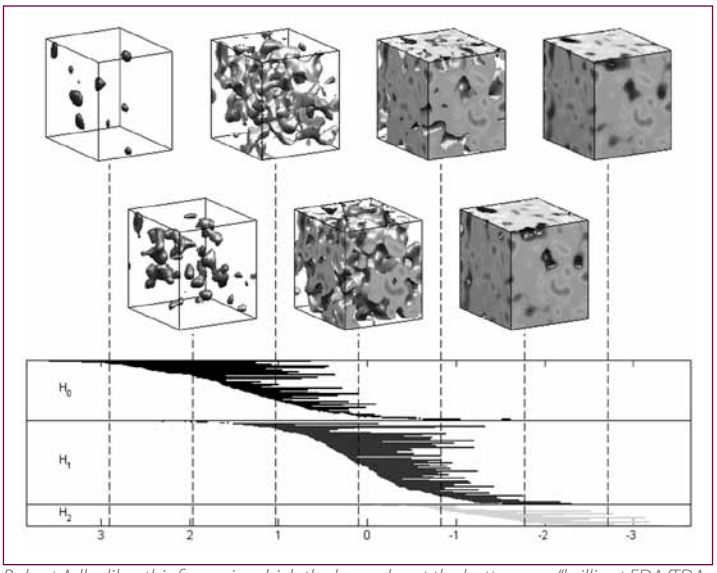
Wikimedia Commons/Ttrung

close to Tukey's plan for EDA, but today's TDA can call on tools of computational topology that were mere pipe dreams in Tukey's day.

In cosmology and astrostatistics, the ideas of TDA have been used for both quantifying the structures behind galactic density data, and for smoothing the data itself. As in manifold learning, the ideas behind TDA-based data smoothing are quite different to the usual ones. Instead of aiming at minimising some quantifiable measure of smoothness such as the L_p norm of a gradient, the aim now becomes to free the data of 'spurious, low level, topology', whatever this might mean.

I plan to devote three more columns to *TOPOS,* developing most of the above thoughts as well as explaining unfamiliar terms and concepts. *En passant,* I will write about my currently favourite figure (right) in which the barcodes at the bottom are brilliant EDA/TDA descriptors of the three dimensional structures at the top.

One column will be devoted to each of the topology, probability, and statistics involved in TDA.



Robert Adler likes this figure, in which the barcodes at the bottom are "brilliant EDA/TDA descriptors of the three-dimensional structures at the top". He will be writing more about this in his next columns.

Today, however, let me conclude by telling you a story, which might help explain why I think statisticians and probabilists should care about TDA.

In 2010 I went to my first Applied Topology workshop. I think I was invited because of work I had done on topology and random fields, but I—like many of you might have been—was very much a statistical–probabilistic fish out of water, gasping in the air of topology, full of terms and ideas that had me floundering. I had a problem. But I am a reasonably quick learner, and it turns out that understanding the basics of algebraic topology—as opposed to breaking new ground in the area—is not all that hard (a point I want to bring home in the next column). So it was not that long until I realized that much of the problem was due not solely to my ignorance, but rather to a community of mathematicians who, to my disbelief, were analyzing data with powerful mathematical tools but with absolutely no use of modern, or even classical statistical methodology to help them. To make matters worse, very few of the speakers had absorbed the most basic statistical concepts that data is often based on a random sample from a larger population, or contains measurement or other errors, so that intrinsic stochastic elements in what they are analyzing could have a major impact on their results.

Although this was less than four years ago, times are a'changing, and a recent IMA workshop on TDA (apparently the best-attended workshop in the IMA's history!) was preceded by a three day tutorial on probability and statistics for topologists. In addition, SAMSI has just run a workshop on TDA as part of its 2013–14 program on *Low-dimensional Structure in High Dimensional Systems,* so not only is the word getting out to applied topologists that they need to think stochastically, the word is reaching statisticians that there is a new application out there that both needs their contribution and is also likely to provide them with new tools that might be the twenty-first century version of EDA.

Topologists and probabilists have also met at an AIM (American Institute of Mathematics) workshop.

So TOPOS is starting to grow. There is no question that the *T* needs *P* and *S*, but there is also no question that the *T* has tools for the *S*, and both have lots of beautiful problems for the *P*'s like me.

More next time. For those who do not want to wait, or who want something more serious than two-page machinations, here are three useful sources that will also direct you further.

- R.J. Adler, O. Bobrowski, M.S. Borman, E. Subag and S. Weinberger, Persistent homology for random fields and complexes. In *Borrowing Strength: Theory Powering Applications, A Festschrift for Lawrence D. Brown*, IMS Collections **6**, 124–143, 2010.
- G. Carlsson. Topology and data. *Bull. Am. Math. Soc. (N.S.)*, **46**(2): 255–308, 2009.
- R. Ghrist. Barcodes: the persistent topology of data. *Bull. Am. Math. Soc. (N.S.),* **45**(1): 61–75, 2008.