Member-only story

Topological Data Analysis (TDA)

A less mathematical introduction

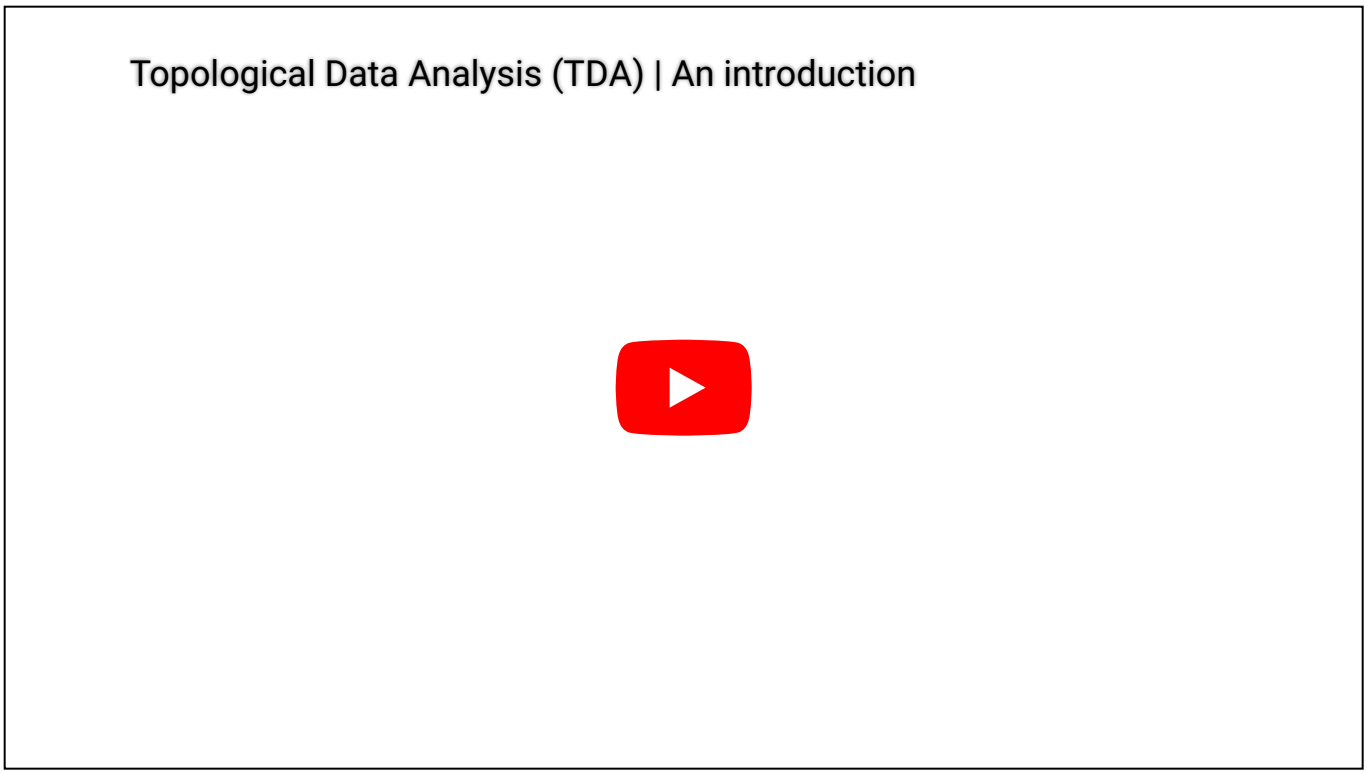
Shaw Talebi

Published in Towards Data Science · 6 min read · May 21, 2022

This is the first article in a series of three on topological data analysis (TDA). TDA is an on-the-rise data science tool that **looks at the *shape* of data.** It consists of various approaches with an underlying theme of extracting structure (i.e. shapes) from unstructured data (i.e. point clouds). In this first article, I will give an accessible introduction to TDA that focuses less on mathematical terminology and more on big-picture ideas. Future posts will discuss two specific techniques under the umbrella of TDA: the **Mapper algorithm** and **persistent homology**.

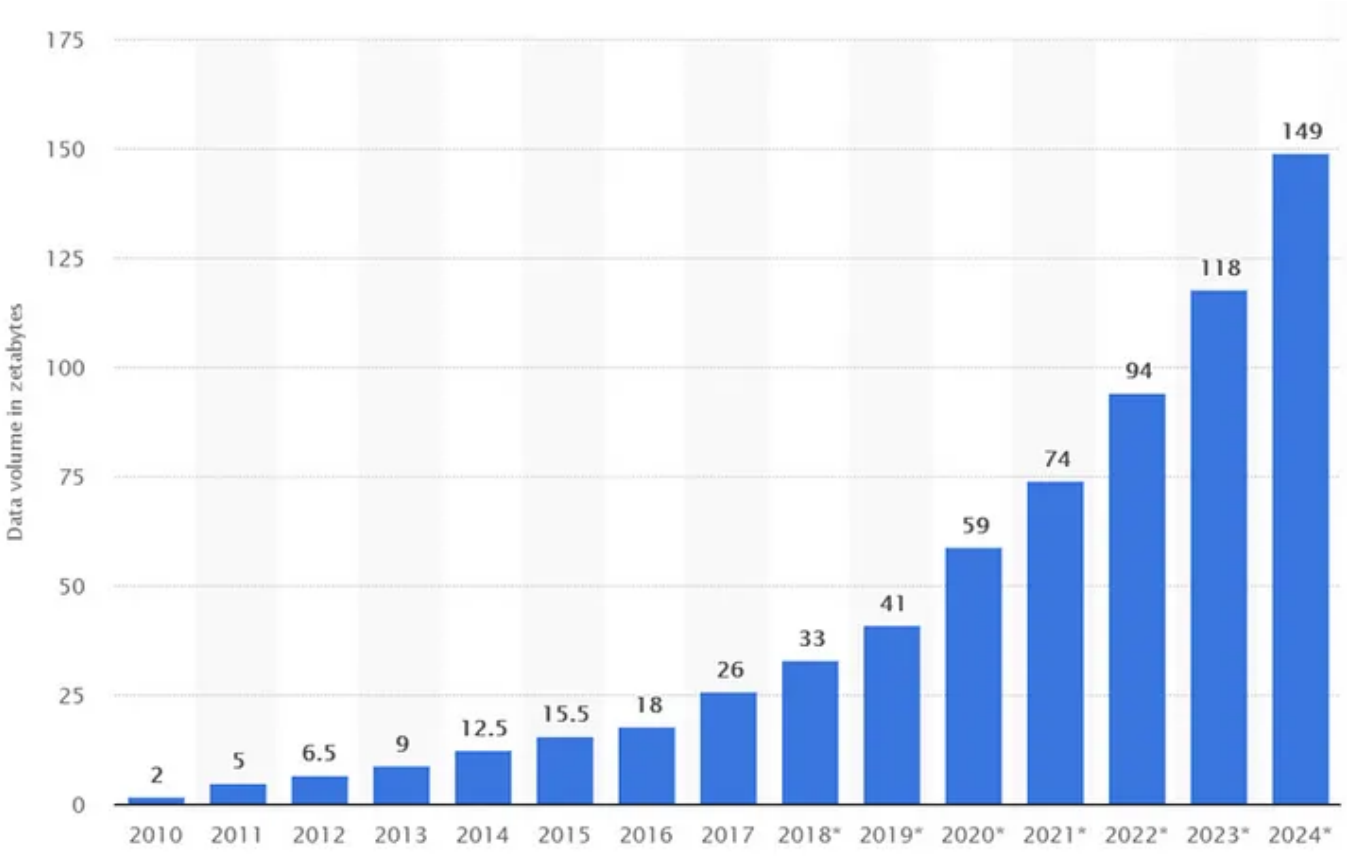
Key points

1. TDA studies the *shape* of data
2. TDA is well-suited for noisy and high-dimensional datasets



The Rise of Data

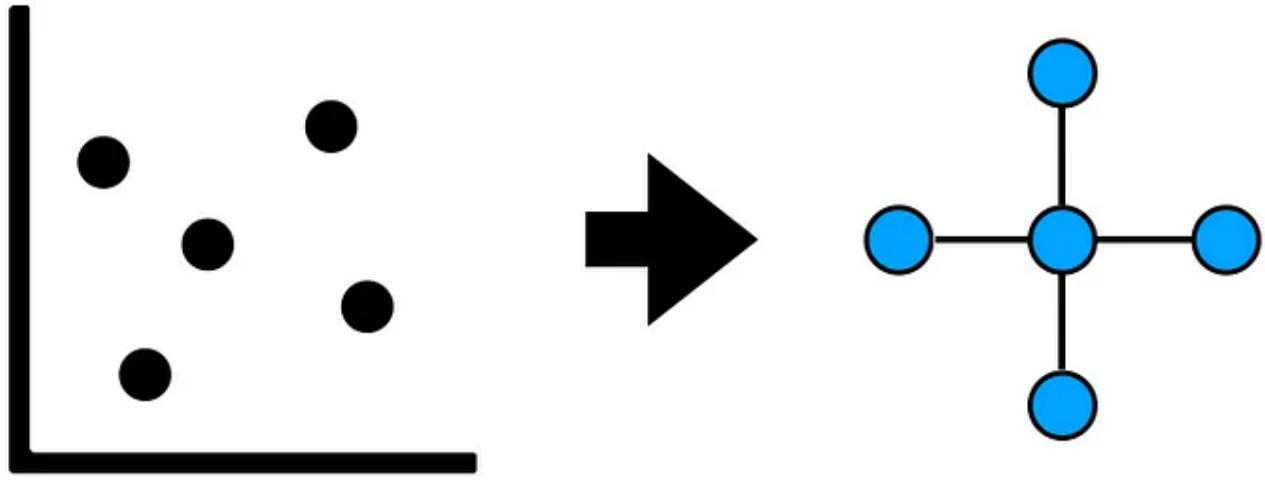
Data volumes across domains seem to be increasing at an accelerating rate. This has served as fuel for many modern technologies e.g. NLP, image recognition, self-driving cars, etc. Although data has been a key for innovation, there's a devil in the details.



The accelerating growth of data volumes. Image source

Mo' data, mo' problems. Any practitioner will tell you, data from the real world is often noisy. A significant amount of care and effort is required to take what we measure and transform it into something we can analyze.

This is where **Topological Data Analysis (TDA)** can help. Rather than working with raw data directly, TDA **aims to extract the underlying shape of data.** From this shape, we can evaluate topological features, which (typically) fare much better to noise than the raw data itself.

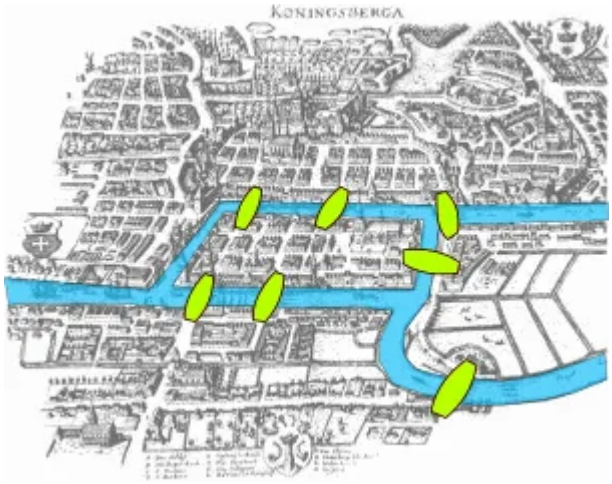


Data → Shape. The basic idea of TDA is to extract shape from data. Image by author.

Topology

TDA is built on ideas from the mathematical field of topology. The story of topology goes back to a famous problem in math called the **7 Bridges of Königsberg**.

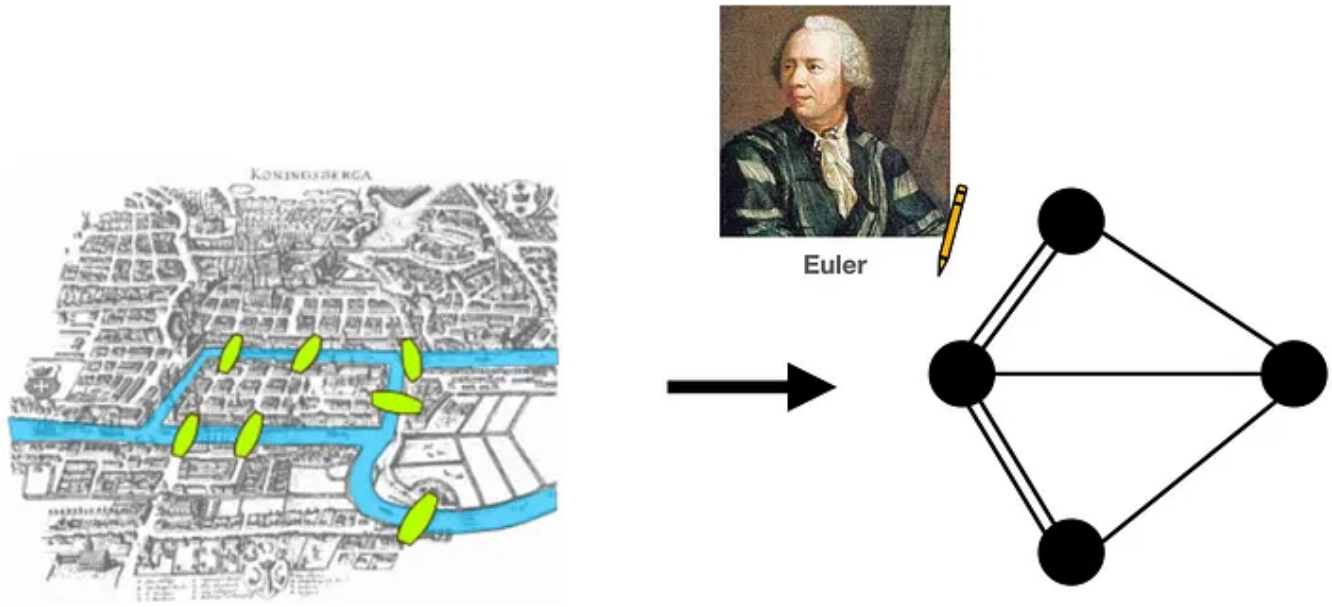
Königsberg was a city consisting of 4 land masses connected together by 7 bridges (pictured below). The famous problem is: *how can one cross all 7 bridges in a single path, but only travel across each bridge once?*



The 7 Bridges of Königsberg (1736). Image source.

If you tried to find such a path with no luck, don't feel bad; no such path exists. However, demonstrating this led famous mathematician Leonhard Euler to lay the foundation for modern-day graph theory and topology.

So, *how did he do it?* What Euler did was draw a picture of Königsberg, but not like the map above, something much more basic. He drew what we now call a **graph**, which is just **a set of dots connected by lines**. The image below illustrates this.



Euler representing the 7 Bridges of Königsberg problem as a graph. Königsberg map from here. Picture of Euler from here.

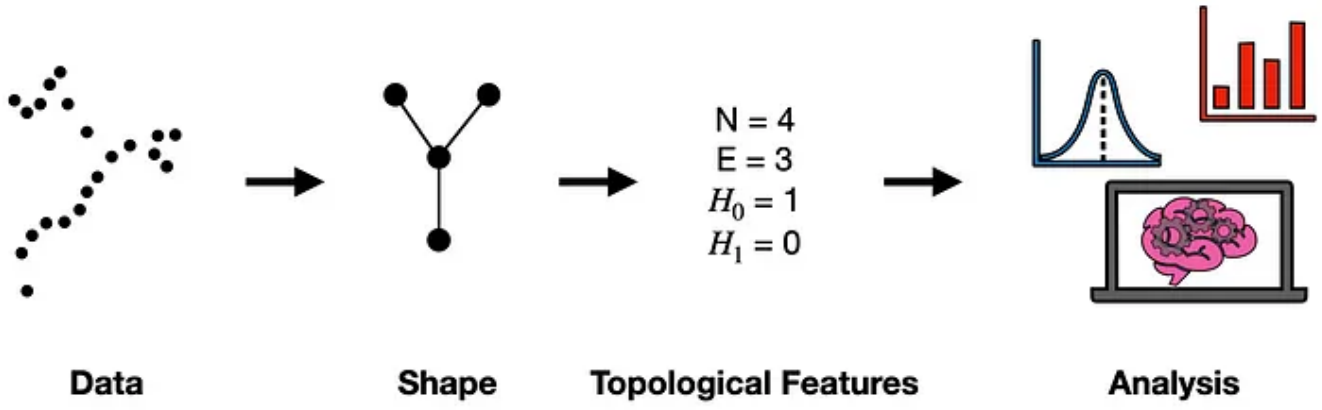
This **graph represents the essential elements of Königsberg** relevant to the problem. Each dot corresponds to a land mass in Königsberg, and two dots are connected by a line if the corresponding land masses are connected by a bridge. This simplified depiction of the problem may seem trivial, but it allowed Euler to solve the problem and change mathematics.

Topological Data Analysis (TDA)

The basic idea of TDA is reminiscent of what Euler did to solve the 7 Bridges of Königsberg problem. We take real-world data (e.g. a map of Königsberg) and extract the underlying shape that is most relevant to our problem (e.g. a graph representing Königsberg). The **key benefit** to doing this is we boil down the data to its **core structure** which is (hopefully) **invariant to noise**.

The Pipeline

TDA isn't a single technique. Rather, it is a collection of approaches with a common theme of extracting shapes from data. A generic TDA pipeline is described in the paper by Chazal and Michel [1]. A visual overview of this pipeline is shown below.



The basic TDA pipeline based on the description by Chazal and Michel [1]. Image by author.

The pipeline starts with data. Generally, we can think of any dataset as being an N-dimensional point cloud. This comes from considering the N variables in the dataset as axes for an N-dimensional space in which data points live.

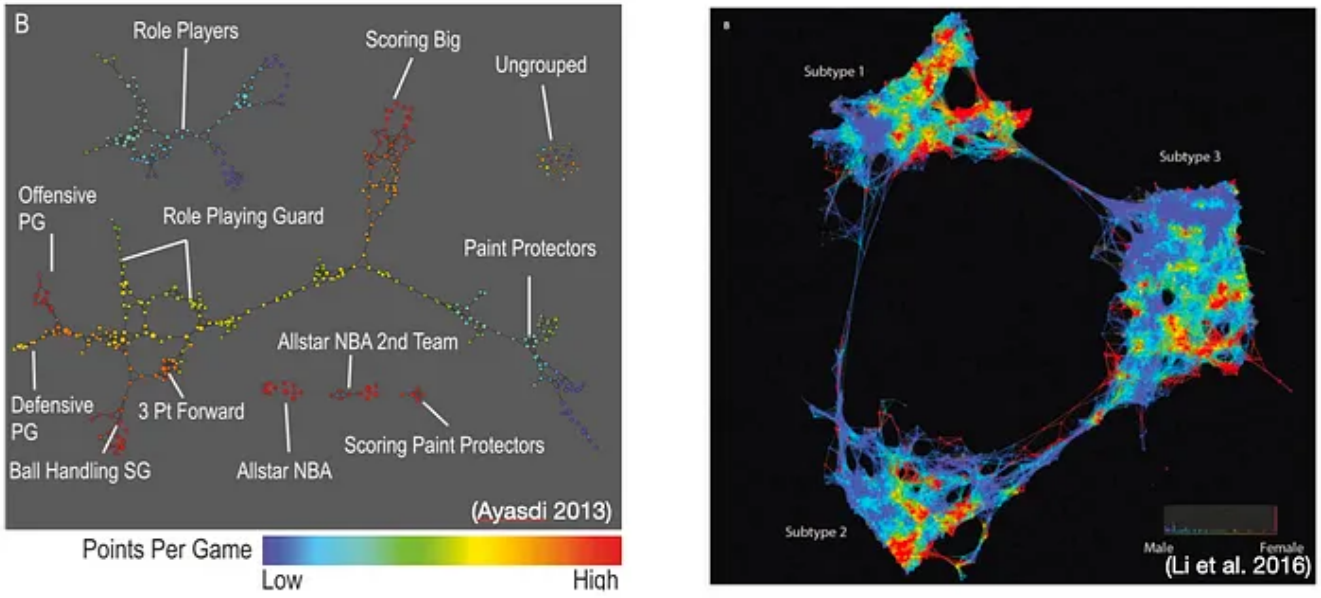
The next step is to generate shapes based on the point cloud. There are many ways to do this, which distinguishes different TDA approaches, as we will discuss in the next blogs in this series.

Once we have the shape of our data, we can characterize its topological features. There are again multiple ways we can characterize shapes. For example, given a graph (like in the image above) we can count the number of dots and lines. We can also turn to something called **homology** and count the **number of "holes"** in the data.

As a final step, we can use these topological features to do an analysis. This could be something as simple as categorizing data based on the statistics of topological features, or something more sophisticated, like using topological features as inputs to a machine learning model.

The Promise

When I was first exposed to TDA, I was really excited. I saw super-cool use cases like redefining basketball positions and discovering type-2 diabetes subgroups. These were presented with engaging visualization, like the ones below, that clearly showed patterns in the data.



Two use cases for TDA. (Left) TDA applied to sports analytics. Image from Ayasdi white paper [2]. (Right) TDA used to identify diabetes subgroups. Image from paper by Li et al. [3].

The Mapper Algorithm

Graph it like Euler

medium.datadriveninvestor.com

The Problem (Nomenclature)

However, the excitement was curbed when I started reading more about TDA. The problem (for me) was most of the literature was heavy on mathematical proofs and jargon. Not to say that this is a bad thing, just that it made it more difficult for me to understand what was going on here. This brings me to the goal of this series.

The Goal

In this series, my goal is to capture the essence of what TDA does without diving too far into the mathematical weeds. I hope this first post has given you a flavor of what TDA is all about. In future posts, I will go into more detail by discussing two specific approaches under the umbrella of TDA: **the Mapper algorithm** and **persistent homology**. Each article will give an introduction to the method and walk through a concrete, real-world example with code.

👉 **More on TDA:** Mapper Algorithm | Persistent Homology

Resources

Connect: My website | Book a call | Ask me anything

Socials: YouTube 📹 | LinkedIn | Twitter

Support: Buy me a coffee ☕

[1] Chazal, F., & Michel, B. (2021). An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists. *Frontiers in Artificial Intelligence*, *4*, 108. https://doi.org/10.3389/FRAI.2021.667963/BIBTEX

[2] Ayasdi (2013). Using Topological Data Analysis for Sports Analytics. http://www.ayasdi.com/wp-content/uploads/_downloads/Redefining_Basketball_Through_Topological_Data_Analysis.pdf (Accessed May 2nd 2022)

[3] Li, L., Cheng, W. Y., Glicksberg, B. S., Gottesman, O., Tamler, R., Chen, R., Bottinger, E. P., & Dudley, J. T. (2015). Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science translational medicine*, *7*(311), 311ra174. https://doi.org/10.1126/scitranslmed.aaa9364

Topological Data Analysis | Topology | Data Science | Machine Learning | Data Analysis