CHARACTERIZATION OF TOPOLOGICAL POINTS IN BIG DATA SETS OF HAUSDORFF SPACES

BY

ONYANGO ALLAN ONYANGO

A Thesis Submitted to the Board of Postgraduate Studies in Partial Fulfilment of the Requirements for the Award of the Degree of Master of Science in Pure Mathematics

SCHOOL OF BIOLOGICAL, PHYSICAL, MATHEMATICS AND ACTUARIAL SCIENCES

JARAMOGI OGINGA ODINGA UNIVERSITY OF SCIENCE AND TECHNOLOGY

Declaration

This thesis is my own work and has not been presented for a degree award in any other institution.

ONYANGO ALLAN ONYANGO

W251/4159/2020

Signature ..... Date .....

This thesis has been submitted for examination with our approval as the university supervisors.

1. Prof. Benard Okelo

Department of Pure and Applied Mathematics

Jaramogi Oginga Odinga University of Science and Technology, Kenya

Signature ..... Date .....

2. Dr. Richard Omollo

Department of Computer Science and Software Engineering

Jaramogi Oginga Odinga University of Science and Technology, Kenya

Signature ..... Date .....

Acknowledgments

This thesis would not have been possible without the extreme expertise, highly professional guidance and unrelenting patience from my supervisors. Much appreciation to Prof. Benard Okelo who shaped my thinking since the early stages of drafting and throughout the study period. Many thanks to Dr. Richard Omollo who answered every question I had about assembling my python code, Artificial Intelligence techniques, and patiently made frequent phone calls to check on my progress and challenges. Many thanks to the members of the Board of Post Graduate Studies through its coordinator, Dr. Erick Okuto, from the Department of Pure and Applied Mathematics, of the School of Biological, Physical, Mathematics, and Actuarial Science, for their positive criticisms and remarkable wealth of wisdom. I thank Mr. Kevin Mugo and Ms. Esther Wambui, who went beyond fact-checking to review my python code. Much gratitude to my late parents for introducing me to critical thinking, and for not being afraid of the mistakes I made while growing up. Mr Maricus Ahomo, while at Ambira High School (Form 3, 2006), you undoubtedly rekindled the hidden mathematics potential within me! I remain indebted. My lovely fiancee, Beryl, your rejuvenating love and peace of mind is all I ever needed. To my sister, Emmah, and brother, Edgar; the support, encouragements and your constant prayers is what I needed most. To my niece Dorine Hills, and nephew Danel, your constant curiosity inspires me every single day. I salute you my great spiritual mentors, Fr. Lawrence Omollo, Mr. Emmanuel Okeyo, and Mr. Edwin Omondi. Finally, I most reverently thank the Almighty God for His great wisdom, perfect health and sufficient grace throughout my study period.

Dedication

To my esteemed late parents, dad, Bathpreston Onyango Oloo, and dear mum, Dorine Atieno, to my lovely fiancee, Beryl, to my admirable sister, Emmah Onyango, and great brother, Edger Onyango, to my beloved niece and nephew, Doreen and Danel.

Abstract

Topological Data Analysis (TDA) is an important aspect in the field of topological data theory since the 21st century's first decade. Modern TDA utilizes the structural characteristics of Big Data (BD), otherwise known as point cloud data sets. Topology and Geometry are tools used to analyze highly complex and multi-dimensional data by creating a summary of these characteristics to uncover hidden features in these datasets, while preserving feature relationships within the data. Describing topological Data Points (TDP) is very intricate due to the nature of BD. This makes it difficult to locate Big Data Sets (BDS) particularly in a general topological space setting. Because of the structure in T_2-spaces, it is even more difficult to locate these BDS in Hausdorff spaces. The objectives of the study include; to characterize TDPs in Hausdorff spaces, to locate BDS in Hausdorff Spaces, and to establish distribution patterns of TDPs in Hausdorff spaces. The methodology involved use of BDS, separation criterion of Hausdorff Spaces, Artificial Intelligence (AI) and Machine Learning (ML) techniques, as well as development of algorithms and simulations using python. The results show that the space of a TDP is compact, and has no less than one closed TDP. Moreover, the set of all condensation points of a TDS is infinite and has infinite cardinality. Lastly, Covid-19 cases are densely distributed in regions experiencing extremely low temperatures. The results of this study are useful to policy makers in the health sector in controlling Covid-19. This work is also a contribution of knowledge in the field of TDA.

Contents

Index of Notations

Chapter 1

INTRODUCTION

1.1 Mathematical Background

TDA is founded under the ubiquitous theory of persistent homology. Some of the pioneer contributors to TDA include Frosini[16], Robins[29] and, Edelsbrunner[13], who founded the notion of how features persist as the data is modified. Nevertheless, the genesis of the term TDA expression appears not to have surfaced till contributions by Carlsson[12], De Silva, and Bremer[2]. Thereafter, Carlsson[5], became instrumental in the popularization of TDA, establishing the ways topological techniques will remedy challenges encountered while implementing topology to analyze BD. Perea [26] put up other developments by observing that Persistent homology is currently one of the more widely known tools from computational topology and topological data analysis.

Topology and Geometry are tools used to investigate highly composite data [7] by creating a compendium of the features of data to uncover hidden attributes within the dataset. Normally, the dataset of interest is often centered around structures that appear challenging to be revealed

with traditional methods[24]. The major TDA approach for removing "topological noise" is to map the original data to a lower dimensional approximation acquired through a multidimentional assortment.

Open sets therefore provide an essential approach to understand nearness of points without a distance element defined in a topological space. Other inherent mathematical concepts to understand besides topology include continuity, connectedness, and closeness, which embrace nearness. The problem is that there isn't a single story happening in this data. We can therefore say this data has much "noise"!

The explosive growth in data, voice and video traffic, and ubiquity of social-media content, health records, and many more data sources, has been a contributing factor of big data. It was anticipated that the generated data volume could be 44 zettabytes in 2020 as found in [3]. Just pose and imagine the billions of emails sent and stored weekly. Using social media to interact and communicate generates immense data quantities as can be revealed through the following statistics [1]:

(i). 350 million photos are uploaded to Facebook daily; close to 100 hourly videos get uploaded to YouTube.com every 60 seconds, while in Instagram, over 45 million pictures are uploaded daily.

(ii). As of 2015, Social Networking sites are used by 72% of online adults, while the percentage of Facebook users not concerned with any kind of privacy control is 25%.

(iii). Approximately 293,000 status updates are posted on Facebook every sixty seconds, while every 48 hours, more than a billion tweets are sent, as daily 1 billion new Twitter accounts created.

(iv). During August 2011 earthquake in Mineral, Virginia, the New York City residents received tweets 30 seconds before they felt it.

The vast data volume with its complexity has propelled technological advancements realized as well as accelerated increase in bandwidth capacity, processing power, storage capability and transfer velocity. This partially, is due to the technological advancement in high power computing.
There is therefore urgent need to establish robust and resilient techniques to process the Big Data. BD consists of 5 Vs: Value, Variety, Volume, Veracity and Velocity[8]. The data size to be processed and analyzed constitutes the volume. The speed of growth and usage of this data is the velocity. The varied data formats cum types is the variety. Veracity involves accuracy plus analysis of the results of the datasets. The richness obtained after the processing the dataset is the value.

The growing volumes of Voice over IP (VoIP), social-media content, [18] underscores the requirement of ways of countering the ambiguity innate the finite datasets. Presently, roughly eighty percent of datasets remains indeterminate. Figure 1.1.1 demonstrates this. TDA has lately recorded advances in innumerable directions and application disciplines. The fundamental aim of TDA is to extract multi-dimensional rich data features based on geometry and topology pre-existing in distributed datapoints as shown in [5] and [25]. Connections within the data and topological methods have a close affiliation to neural networks between data-points which reveal insight into this united structure. According to [15], most commonly, every other form of TDA revolves around the steps below:
Firstly, the data sample is presumed as datapoints which are finite quan-

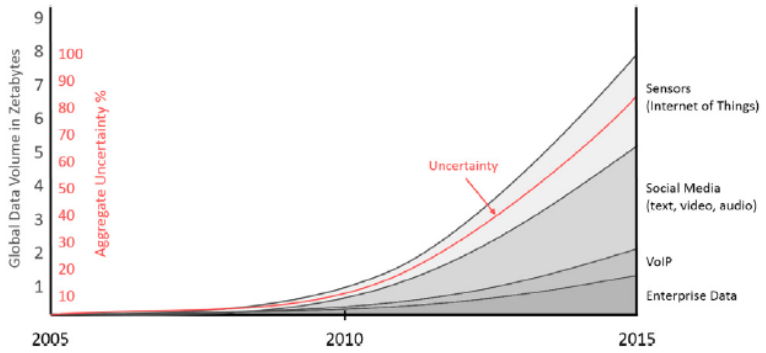

Figure 1.1.1: Predicted uncertainty of Big Data [32].

tified as a metric space $\mathbf{R}^d$. Worth to mention is that the metric choice may be vital to guarantee remarkable topological and geometric features of the data set. Secondly, to tap more from the fundamental concepts of Geometry and Topology, a mathematical structure is computed on top of the dataset. This is mostly cases in SC or a convention of SCs, that depicts the high dimensional data structures at varied degrees.

Thirdly, from these high dimensional data structures built on atop the data set, topological or geometric information is derived. The shape of the data from which we extract the topological/geometrical high dimensional features can either be crude structural summaries or relevant approximations that need further approaches like persistent homology and visualization. The extracted topological and geometric information gives rise to insightful features and descriptors into the data which when injected into further analysis and machine learning procedures, reveal very rich results and significant meaning that can be used in other disciplines like medicine, biology and astrophysics, just to mention a few.

1.2 Basic Concepts

Here, we cover the basic concepts key to our study on characterization of topological points in Big Data sets of Hausdorff spaces (HS).

Definition 1.1. ([31]Definition 1.1.1) **Topological space**
Let $\mathbf{X}$ be a non-empty set and τ a collection of all subsets of $\mathbf{X}$. Then τ is a topology on $\mathbf{X}$ if the following axioms are satisfied:

(i). $\emptyset, \mathbf{X} \in \tau$

(ii). Any arbitrary union of sets belonging to τ also belongs to τ.

(iii). Any finite intersection of sets belonging to τ also belongs to τ.

Definition 1.2. ([31]Definition 6.1.1) **Metric Space**
Let $\mathbb{X}$ be a non-void set and $\rho : \mathbf{X}$ x $\mathbf{X} \to \mathbb{R}$ be a non-negative function satisfying the following properties;

(i). $\rho(x, y) = \rho(y, x) \forall x, y \in X$ (Commutativity property/symmetry axiom). The distance from one point to another is the same as going in the opposite direction.

(ii). $\rho(x, y) = 0$ if and only if $x = y$ (Zero Property or Non-negativity axiom) The distance between two points is a positive number, and the distance is 0 if and only if the points are the same.

(iii). $\rho(x, z) \leq \rho(x, y) + \rho(y, z) \forall x, y, z \in X$ (Triangle Property). The distance between two points x and z is always no longer than taking a detour through point y.

Then $(\mathbf{X}, \rho)$ is called a metric space and the function ρ is called a metric on the set $\mathbf{X}$. The elements of $\mathbf{X}$ are then called the points of $\mathbf{X}$.

Definition 1.3. ([10]) **Data**

Data refers to facts, quantities, characters, symbols gathered for investigation and analysis.

Definition 1.4. ([33]) **Data Set**

A composition of related information consisting of unconnected datapoints stored, ordered and viewed as a component.

Definition 1.5. ([19]) **Big Data**

Big Data can be described as hugely large, highly composite datasets to be analyzed by traditional techniques, but might reveal structure, patterns, relationships, shapes, when computational analysis methods are involved.

Definition 1.6. ([22]) **Artificial Intelligence (AI)**

Artificial Intelligence (AI) refers to an extensive branch of computer science involved in the theory and development of smart computer systems having the capability to perform tasks that usually require human intelligence. e.g. natural languages translation, voice recognition and visual perception.

Definition 1.7. ([22]) **Machine Learning (ML)**

Machine Learning (ML) refers to a discipline of AI that utilizes computer algorithms to learn, experience, adapt and automatically improve without human programming.

Definition 1.8. ([27]) **Python**

Python is a high-level, object-oriented, interpreted, general-purpose pro-

gramming language for prototyping, website coding, application development, processing images and scientific data analysis.

Definition 1.9. ([31]) $\mathbf{T_0 - Space}$

Let (X, τ) be a topological space. Let x and y be two distinct points in $\mathbf{X}$. There exists an open set $\mathbf{G} \in \tau$ which contains x and not y i.e. $x \in \mathbf{G}$ and $y \notin \mathbf{G}$.

Definition 1.10. ([31]) $\mathbf{T_1 - Space}$

Let (X, τ) be a topological space. Let x and y be two distinct points in $\mathbf{X}$, i.e. $x \neq y$. Let $\mathbf{U}$ and $\mathbf{V} \in \mathbf{X}$ be two open sets such that $x \in \mathbf{U}$, $y \notin \mathbf{U}$ and $y \in V, x \notin \mathbf{V}$.

Definition 1.11. ([31]Definition 6.1.24) $\mathbf{T_2 - Space}$

Let $(\mathbf{X}, \tau)$ be a topological space. Let x, y be two distinct points in $\mathbf{X}$. i.e. $x \neq y$. Let $\mathbf{U}, \mathbf{V}$ be two open sets of $\mathbf{X}$ such that $x \in \mathbf{U}, y \in V$ and $\mathbf{U} \cap \mathbf{V} = \emptyset$.

The topological space $(\mathbf{X}, \tau)$ that satisfies the $\mathbf{T_2 -}$ *space* is referred to as Hausdorff Space.

Example 1.12. Let $\mathbf{X} = \{1, 2, 3\}$ and $\tau = \{\emptyset, \{2\}, \{2, 3\}, \{3\}, \{1, 2\}, \{1, 3\}, \mathbf{X}\}$. $x = 1$, $y = 2$ so, $x \neq y$. $\mathbf{U} = \{1\}$ and $\mathbf{V} = \{2, 3\}$ then $\mathbf{U} \cap \mathbf{V} = \emptyset$.

Definition 1.13. Condensation Point

Let $\mathbf{X}$ be a topological space and $\mathbf{S}$ a subset of $\mathbf{X}$. Then the condensation point $\mathbf{P} \in \mathbf{S}$ is any point $\mathbf{P}$ such that every neighborhood of $\mathbf{P}$ contains uncountably many elements of $\mathbf{S}$.

Definition 1.14. Hausdorff Maximal Principle (Zorn's Lemma)

Let $\mathbf{X}$ be a partially ordered set, and let every linearly ordered subset of $\mathbf{X}$ have an upper bound. Therefore, $\mathbf{X}$ contains a maximal subset.

Definition 1.15. ([31]Definition 10.3.27) **Regular Topological Space**
Let $(\mathbf{X}, \tau)$ be a topological space. Let $\mathbf{F}$ be a closed subset of τ. Let x be a a distinct point in $\mathbf{X}$ such that x is not in $\mathbf{F}$. Let $\mathbf{U}, \mathbf{V}$ be two disjoint open sets, one containing $\mathbf{F}$ and the other containing x, such that $\mathbf{F} \subset \mathbf{U}, x \in \mathbf{V}$. $\mathbf{U} \cap \mathbf{V} = \emptyset$. Therefore, $(\mathbf{X}, \tau)$ is a regular topological space.

Definition 1.16. ([31]Definition 10.3.27) $\mathbf{T}_3 - \mathbf{Space}$
A regular topological space $(\mathbf{X}, \tau)$ which is also $\mathbf{T}_1$-space is a $\mathbf{T}_3$-space.

Definition 1.17. ([31]Definition 10.3.20) **Normal Topological Space (NTS)**
Let $(\mathbf{X}, \tau)$ be a topological space and $\mathbf{F}_1, \mathbf{F}_2$ be two disjoint closed subsets of $\mathbf{X}$. Then there exists two disjoint open sets $\mathbf{U}, \mathbf{V}$ of $\mathbf{X}$ such that $\mathbf{F}_1 \subset \mathbf{U}$ and $\mathbf{F}_2 \subset \mathbf{V}$. Therefore, $(\mathbf{X}, \tau)$ is a normal topological space.

Definition 1.18. ([31]Definition 10.3.20) $\mathbf{T}_4 - \mathbf{Space}$
A $\mathbf{T}_4$ Space is a normal space which is also $\mathbf{T}_1$-space.

Definition 1.19. ([21]Definition 3) **Open set**
A subset $\mathbf{A} \subset \mathbf{X}$ of the topological space $\mathbf{X}$ is an open set of $\mathbf{X}$ if it belongs to τ.

The following are properties of open sets:

(i). The empty-set $\emptyset$ is open.

(ii). Any arbitrary union of open sets is open.

(iii). Any finite intersection of open sets is open.

Definition 1.20. ([21]Definition 4) **Closed set**

The subset $\mathbf{A} \subset \mathbf{X}$ of the topological space $\mathbf{X}$ is a closed set of $\mathbf{X}$ if its complement $\mathbf{X} \backslash \mathbf{A}$ is open.

Definition 1.21. ([21]Definition 6) **Compact topological space**

A topological space $\mathbf{X}$ is compact if every open covering of it contains a finite sub-collection that is also a covering of $\mathbf{X}$.

Definition 1.22. ([21]Definition 28) **Connected topological space**

A topological space $\mathbf{X}$ is connected if for any two points of $\mathbf{X}$ there exists a path between them on $\mathbf{X}$.

Definition 1.23. ([21]Definition 29) **Connected components**

The maximally connected subsets of a topological space $\mathbf{X}$ are called its connected components.

Definition 1.24. Topological Data Point

Letting $\mathbf{H}$ to be a nonempty compact Hausdorff space, a point $a \in \mathbf{H}$ is called a topological data point (TDP) if $\mathbf{H} \setminus \{a\}$ is a compact subspace of $\mathbf{H}$.

Definition 1.25. TDP Space

A nonempty compact Hausdorff space $\mathbf{H}$ is called a TDP space if every $a \in \mathbf{H}$ is a TDP.

Remark 1.26. Letting $\mathbf{H}$ be a topological space, then $\mathbf{H} = \mathbf{P} \dagger \mathbf{Q}$ means $\mathbf{P}$ and $\mathbf{Q}$ are nonempty subsets of $\mathbf{H}$ such that $\mathbf{H} = \mathbf{P} \cup \mathbf{Q}$ and $\mathbf{P} \cap \overline{\mathbf{Q}} = \overline{\mathbf{P}} \cap \mathbf{Q} = \emptyset$.

1.3 Statement of the problem

Describing topological points is very intricate due to the nature of Big Data. This makes it difficult to locate Big Data sets particularly in a general topological space setting[9]. Because of the structure in T_2-space, it is even more difficult to locate these Big Data sets in Hausdorff spaces. In spite of the remarkable efforts put up by traditional techniques in data analysis, these techniques have not always kept up with the exploding data quantity and complexity since the techniques often rely on overly simplistic assumptions and approximations in their computations[14]. Besides, these techniques do not pay attention to the arbitrariness of data and the inherent unpredictability of the datasets. Accordingly, these techniques are exploratory, lacking the efficiency to distinguish information of interest from "topological noise". The Vietoris-Rips complex for a parameter t has been so ubiquitously used to build a useful simplicial complex to mirror the data structure and utilizes the original data as the vertex set. The bone of contention, however, has always been how to choose the t parameter such that the Rips complex reveals the structure of the underlying data set. It is precisely this question that appeals to our conscience of thought towards a Hausdorff Space as a topological signature of the data set. Finally, previous studies done on Topological Data Analysis have had very little focus on the application of Hausdorff spaces.

1.4 Objectives of the study

Below is the main objective of our study:

(i). To characterize topological points in Big Data sets of Hausdorff Spaces.

Below are the specific objectives of our study;

(i). To characterize topological data points in Hausdorff spaces.

(ii). To locate Big Data sets in Hausdorff Spaces.

(iii). To establish distribution patterns of topological data points in Hausdorff spaces.

1.5 Significance of the study

Big Data sets contain millions of multi-dimensionally rich features and hundreds of thousands, of measurements. These features when analyzed, become extremely significant in areas such as computer vision domains, image recognition, computational finance. Natural language processing, smart keyboards and automatic email reply suggestion, machine translation, spelling correction. In transportation, self-driven cars, health care implementations such as digital disease detection, disease prediction patterns and many more. In Education, BD and DL will contribute to improvement of the competency of systems of education as well as output from the students. As the variety of data being generated expands and speed of data generated skyrockets, the emergence of new technologies like Internet of Things (IoT), Automation, mobile technology and cloud computing, analyzing Big Data data will provide overwhelming opportunities[20].

Chapter 2

LITERATURE REVIEW

2.1 Introduction

We describe literature in Topological Data Analysis using techniques like Persistent Homology and Simplicial Complex. We then explore BDS with the extremely large and highly complex properties. We thereafter discuss the characterization and significance of the Hausdorff Spaces. Next, we describe simulations within the python programming language with applications in TDA. Finally, we establish a research gap summary of TDA from the beginning to the latest [20].

2.2 Topological Data Analysis

The principle idea behind TDA involves the application of techniques in recognition of shape and patterns within data. The finite TDPs within a Euclidean space becomes the driving force for Big Data to be considered in TDA. TDA perceives the point cloud data as a discrete cluster of points

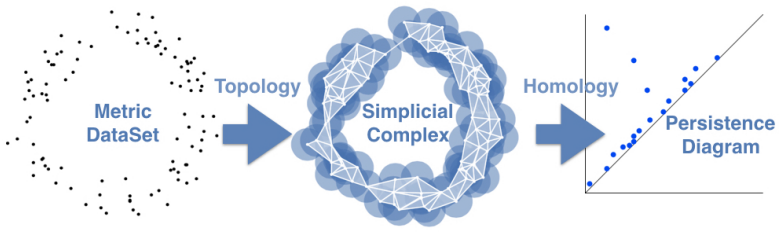

Figure 2.1.1: Encoding pointcloud persistent topological features (left) by approximation of the space by simplicial complexes (middle) into a persistence diagram (right)[20].

within a compact topological space infinitely composed of many points. When neighboring data points are "connected" to reveal geometry atop the dataset, this reveals rich topological features. The idea of distance closeness among TDPs is the backbone that TDA exploits to qualify data sets as metric spaces. The driving force behind Topological Data Analysis is to construct higher-dimensional structures by linking pairs of TDPs by edges as well as by $(k+1) - tuple$ of surrounding TDPs. This drives us to a concept called simplicial complexes, which makes it easy to recognize emerging topological characteristics including points, lines, holes, cycles, and voids.

2.2.1 Persistent Homology

Given that $u \geq v$ then non-empty sets $A_u \subseteq A_v$ such that u and v are distinct points in the set A. Moving from u to v, the components of non-empty set A_u may merge as new ones are born, which have higher chances of merging with each other or with the existent components of A_u. Consequently, these components may change in their topology with

holes and other structures forming and disappearing. This is a perfect demonstration of persistent homology. The 'persistence' idealogy is as a result of the changing of the level u with no change in the homology until a critical point f of level u is reached; which means that the topology of the excursion sets 'persists' (remain static), between the varying heights of critical points. This concept of persistent homology (PH) persists further; such that every time two components merge, we treat the first of these to have appeared as though its existence continues beyond the point of merging.

Moving forward, a more promising illustration of persistent homology is through barcodes. Suppose $dim(M) = N$, and given a smooth of f, if A_u is nonempty, then $dim(A_u)$ seamlessly becomes N. The barcode for the excursion set f becomes the collection of $N + 1$ graphs, with each homology group having one. A bar in the $k - th$ graph, beginning at u_1 and terminating at $u_2(u_1 \geq u_2)$ reveals $H_k(A_u)$ that emerged at u_1 and vanished at u_2.

Figure 2.2.2 displays a more impressive illustration with a three dimensional view. It is imperative to indicate that unlike a two dimensional (2D) space, our comprehension is served with immense visual information from the barcodes since the increase in the N-dimension parameter space projects more clarity. As a result therefore, it becomes uncomplicated to observe barcodes with six sets of bars for the six persistent homologies [28].

The outputs above depend on the assumption that the squared distance function d_K^2 is 1-semiconcave. This means that the function $x \rightarrow \|x\|^2 - d_K^2(x)$.

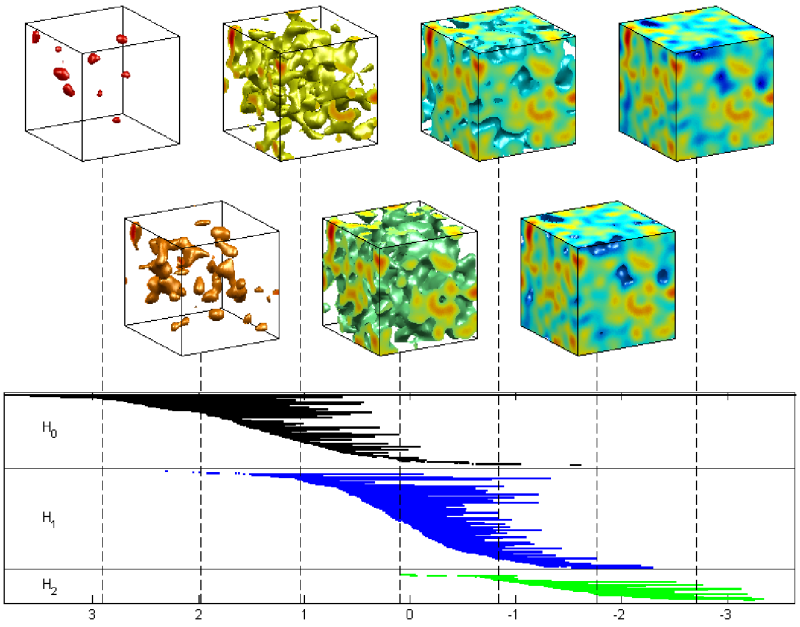

Figure 2.2.1: Bar-codes from the expression of sets within 3-D random space. The top seven boxes reveal the sets and field values decorated in colors. [28].

2.3 Big Data Sets

Big Data Sets can be described as hugely large, highly composite datasets to be analyzed by traditional techniques, but might reveal structure, patterns, relationships, shapes, when computational analysis methods are involved[22]. BD is found almost everywhere with many connections and meaning hidden within this complex data. Precisely, five "Vs" (Value, Veracity, Volume, Velocity and Variety) are used to sum up the description of BD. Given the traditional software tools, every attempt to store, extract, search, share, analyze or process this massive and complex data set has achieved little success. Despite this failure, different levels of society, including businesses, Education, health, transport, research and

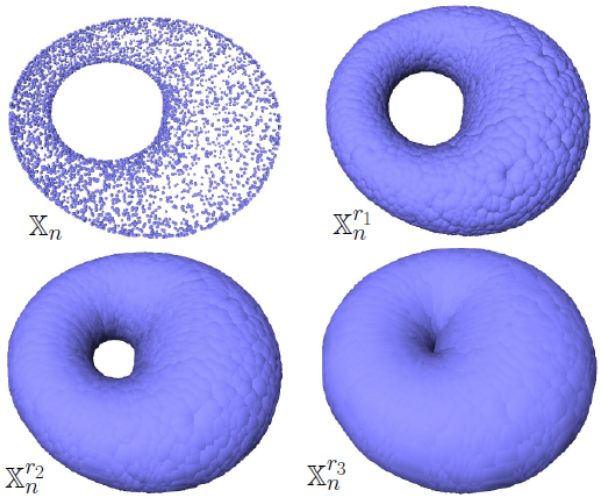

Figure 2.2.2: The point cloud instance $\mathbf{X}_n$ derived from the torus surface ($\mathbf{R}^3$ in top left) which leads to varying radii $r_1 < r_2 < r_3$ [15].

many more continue to persistently and desperately sort insights from Big Data analysis to enhance their efficiency and performance. Human ability to visualize Big Data Sets is not keeping pace with its exploding nature of production. It is therefore urgent to come up with advanced and proficient analysis methods that will manage this kind of data. One of such methods is the application of Geometrical and topological tools. Geometry involves the study of distance functions which works very well with large finite data sets. The mathematical concepts that were formulated by mathematicians, that unite both topological and geometric techniques often revolve around point clouds, which are basically finite point sets that are equipped with a distance function. An important objective in BD analysis is to understand how the data is organized in large scale hence retrieving qualitative information about the data. For instance, given a data set of diabetic patients, insightfully distinguishing the two

distinct forms of the disease is firstly important. Given a very large data set $\mathbf{X}$, it may prove challenging to apply a clustering algorithm on top of the data set. One may otherwise opt to cluster subsamples from $\mathbf{X}$. The confronting question is always whether sampling a cluster is enough proof of representation of the whole data set. An alternate approach is to construct two clusters from the whole dataset, assuming confidence in their consistence. Hence, taking the subsamples $\mathbf{X}_1$and $\mathbf{X}_2$, including their union $\mathbf{X}_1 \cup \mathbf{X}_2$. A clustering scheme is then applied to each of the sets separately by denoting the three sets $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_1 \cup \mathbf{X}_2$ by sets of clusters $C(\mathbf{X}_1)$, $C(\mathbf{X}_2)$, $C(\mathbf{X}_1 \cup \mathbf{X}_2)$ respectively. Suppose the clustering scheme on the data sets induced maps of the collection of clusters, i.e. functorial, then a diagram of the sets would be derived as follows in figure 2.3.1.

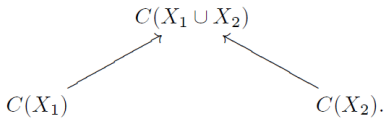

Figure 2.3.1:

If these clusters in $C(\mathbf{X}_1)$ and $C(\mathbf{X}_2)$ in $C(\mathbf{X}_1 \cup \mathbf{X}_2)$ consistently correspond under the maps, then its enough to deduce that subsample clustering corresponds to the clusterings on the entire data set $\mathbf{X}$. Elsewhere, given a varying big data set $\mathbf{X}$, clusters can appear, disappear, merge or even split in distinct clusters. Functoriality can be instrumental in studying this analysis behavior. For all $t_0 < t_1$, we represent TDS within clusters t_0 and t_1 as $\mathbf{X}[t_0, t_1]$. For all $t_0 < t_1 < t_2 < t_3$, the point cloud data set results in a diagram 2.3.2. If functoriality is applied in the clus-

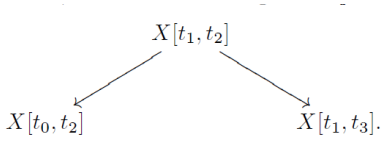

Figure 2.3.2:

tering scheme, a correspondent diagram 2.3.3 of the data set is obtained. This set will contain an over time clustering behavior as revealed in the

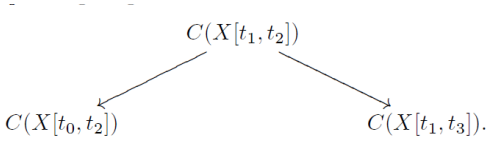

Figure 2.3.3:

illustration. The illustration 2.3.4 corresponds to a unit cluster at an initial time t_0, which breaks into two collections within the interval $[t_1, t_2]$, that finally merges back to the interval $[t_2, t_3]$. Despite the development

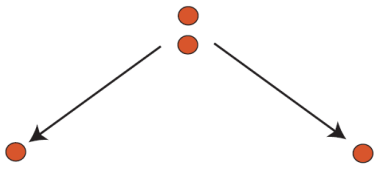

Figure 2.3.4:

of fresh BD analysis techniques for complex BDS, shape identification and interpretation has increasingly proven more challenging to visualize. Because of the existence of more structure to be mined from this BDS than the traditional ones can output, a remarkable new method of "shape" identification of these BDS is TDA.

Several methods exist to construct shape from point cloud data. One of

these methods is described herein: We encircle every TDP within a "ball" whose radius is ε centered within the TDP. While ε increases in size, the cluster no longer looks like isolated points, but gradually gains shape. As it gets larger, an irregular unit component emerges. This technique is therefore used to generate a SC, which begins with vertices as TDPs. Wherever intersection occurs between any two balls, an edge is inserted in the middle, while as intersection among any 3 balls occurs, a three edges bounded face is added. As this process continues, a high-dimensional n-face with n+1 intersections is created. This is referred to as a $\check{C}ech$ complex.

2.4 Topological Data Points

A topological data point (TDP) $a \in \mathbf{H}$ refers to a compact subspace of $\mathbf{H}$ if $\mathbf{H} \setminus \{a\}$ is a nonempty compact Hausdorff space. In Euclidean spaces, the idea of closeness, or limits of points can be described with reference to relationships between sets rather than distance. A topological space refers a set of points, together with the neighborhoods around each set of point, that satisfy a the axioms of the neighborhoods around each set points.

2.4.1 Connectivity of the Data

An open ball is a ball around a point x_0 of radius r within $\mathbf{X}$ expressed as $\mathcal{B}_d(x_0, r) = \{y | d(x, y) < r\}$ and this implies the topology of a distance function d induced around a radius of a point $\mathbf{X}$ with it's boundary

excluded. Given a set $\mathbf{X}$, a simplex is the complete graph induced on the set. The metric topology within the Euclidean plane is said to be connected. Connectedness of a topological space is a clear indication of non-existence of separation within a space. Separation is the existence of nonempty, disjoint subsets $\mathbf{U}$ and $\mathbf{V}$ of a set $\mathbf{X}$ induced with a topology. Taking $\mathbf{X}$ and draw a ball $\mathcal{B}(x_i, r)$ around each point x_i for some small initial radius r (figure 2.4.1). As r increases at intervals, we can establish how $\mathbf{X}$ connects at different radii by observing different snapshots [11].

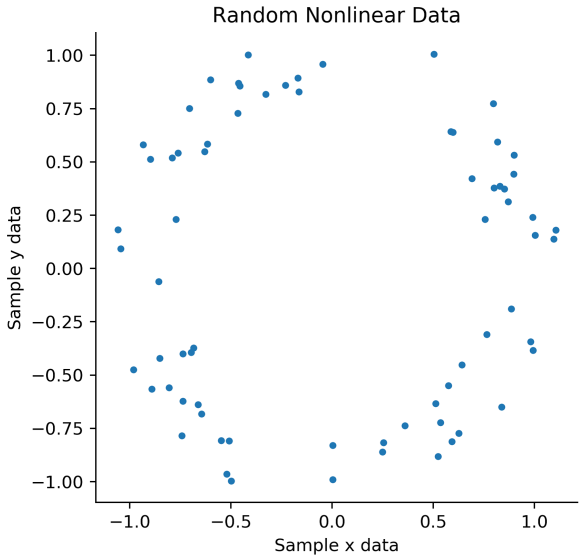

Figure 2.4.1: The sample $\mathbf{X}$ of data [11].

At this point, we will keenly observe the initial intersection of the closures of $\mathcal{B}(x_i, r)$ and $\mathcal{B}(x_j, r)$ on a point within the plane. Without loss of generality, we can confidently state that x_i and x_j are $2r$ distanced apart. Connecting them with an edge, we get a visual graph(figure 2.4.2).

Initially, we observe sparse clusters of edges connecting vertices. As r increases, more edges emerge in small connected clusters for a given

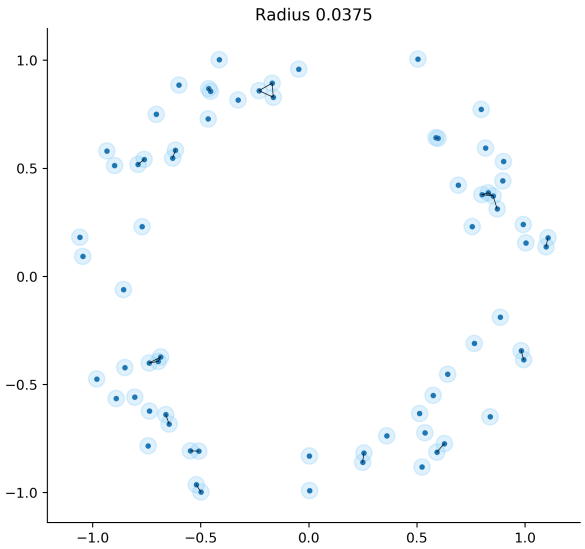

Figure 2.4.2: Initial radius [11].

radius as shown in figure 2.4.3.

We now begin to see the first instance of structure on the data which we can call the connectivity information of the data. These clusters reveal valuable information about the underlying features such that when some initial conditions of the experiment are fulfilled, the data points are more likely to assemble in a certain manner(figure 2.4.4).

Thereafter, notable clusters may begin to emerge, varying from those achieved with smaller radii, basically encompassing them. As the radius of the balls increase, a single component of $\mathbf{X}$ is finally observed for the very first time. The component can among other things be a loop, a chain, have holes or loops, have multiple flares, or another structure(figure 2.4.5).

With the further increase of the radius r, the loop in the data begins to emerge as seen in figure 2.4.5. So as the radius r increases towards r^*, the balls around the points x_i and x_j intersect and the final edge emerges

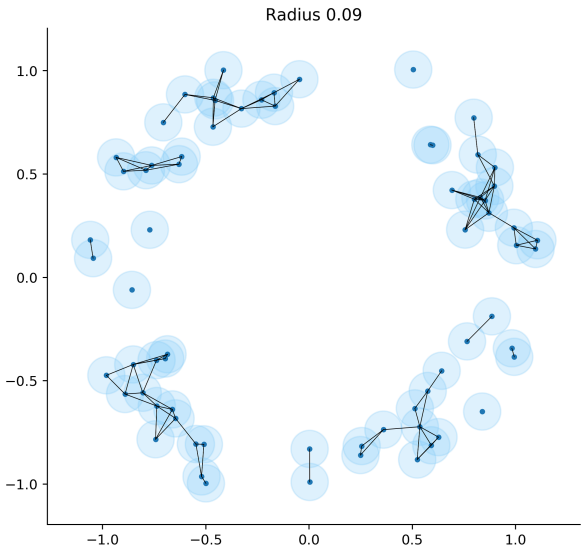

Figure 2.4.3: Denser clusters due to increasing connections [11].

to complete the loop. When eventually this radius is attained, further increment of the radius reveals no further structure for a significant time period. We therefore conclude that the ranges of the radius is $r > r^*$ as shown in Figure 2.4.5 and figure 2.4.6.

2.5 Hausdorff Spaces

The Hausdorff distance was named after the German mathematician, named Felix Hausdorff. To understand **(Hausdorff Spaces)**[31] we first define a $\mathbf{T}_2 - axiom$. Let $(\mathbf{X}, \tau)$ be topological space. Let $x, y \in \mathbf{X}$ be two distinct points such that $x \neq y$. Let $\mathbf{U}, \mathbf{V} \in \mathbf{X}$ be two open sets such that $x \in \mathbf{U}, y \in V$ and $\mathbf{U} \cap \mathbf{V} = \phi$. The topological space $(\mathbf{X}, \tau)$ that satisfies the $\mathbf{T}_2 - axiom$ is called a Hausdorff Space. Describing topological points is very intricate due to the nature of Big Data. This makes it difficult to

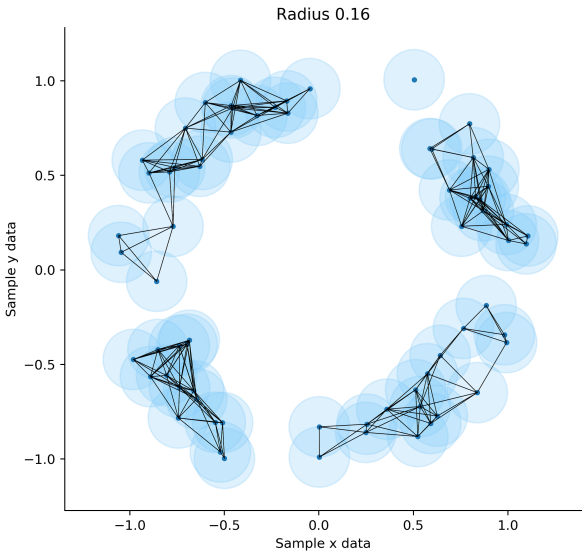

Figure 2.4.4: Clusters in $\mathbf{X}$ [11].

locate Big Data sets particularly in a general topological space setting[9]. Because of the structure in T_2-space, it is even more difficult to locate these Big Data sets in Hausdorff spaces.

2.6 Simulations in Python

Python is popular programming language that is widely used in Artificial Intelligence communities. Python possesses simplicity and readability features for its syntax and this reduces the time taken to test complicated algorithms through minimal code in comparison to the existing languages[23]. Python prides itself with a great numbers of rich library modules for Machine Learning. Furthermore, python commands overwhelming community of developers globally, who generously share troubleshooting and debugging tips through online platforms. The ubiq-

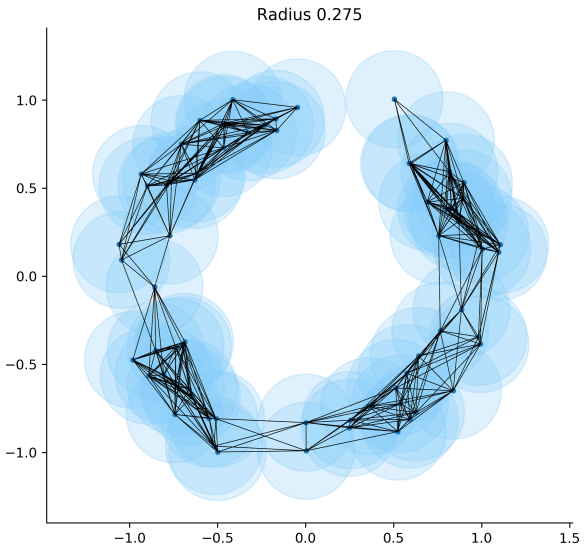

Figure 2.4.5: Largest singular component of $\mathbf{X}$ [11].

uitous nature of Python has granted its usage in nearly all research institutions as well as commercial applications of Deep Learning and Machine Learning. Mapper is a python algorithm that works by constructing a simplicial complex (or graph) from a data set by leveraging a projection function to perform local clustering. As a result, this reveals the topological features of the space. Mapper therefore is an unsupervised TDA technique through which a visual representation of the data set is generated, which so reveals new insights of the data sets, that traditional analysis techniques cannot reveal. Kepler Mapper is a python library for implementing the Mapper algorithm. KeplerMapper is often used for visualization of high-dimensional and 3D point cloud data sets. Every language, python not excluded, has its own limitations. Being an interpreted language that executes code line by line, python keeps it slower as compared to C or C++. Python is the best server-side coding language. When it comes to mobile development, Python is not very good. In case

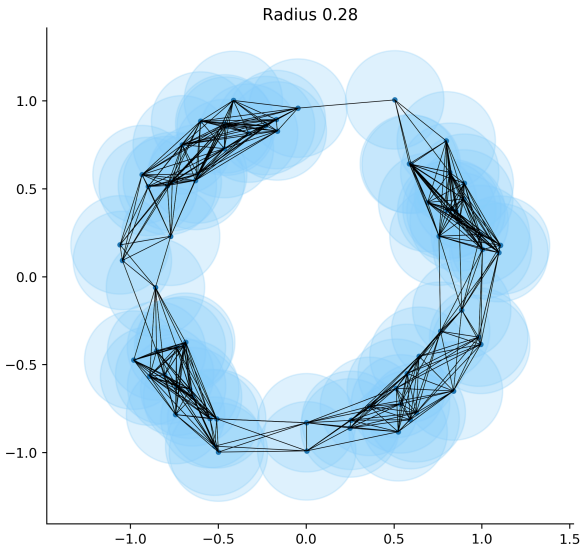

Figure 2.4.6: Before increasing r and first obtaining the loop structure for some $r*$ such that 0.275 <r*< 0.28 [11].

of memory limitation in your project then Python may be bad news to use since its memory consumption is higher. Python being a dynamically typed language, you don't need to mention data type within programs which may end up with run time errors.

2.6.1 DyNeuSR

According to [4], despite the existence of various Mapper software that construct such shape graphs for different kinds of data sets, none of them has an in-built user interface to explicitly support analysis and visualization in neuroscience (see Figure 2.6.1). DyNeuSR is an open-source module used for implementing interactive visualization and neuroscientific computation of topological structures. DyNeuSR has also been useful in the estimation of variation in the activities of the brain. From Fig-

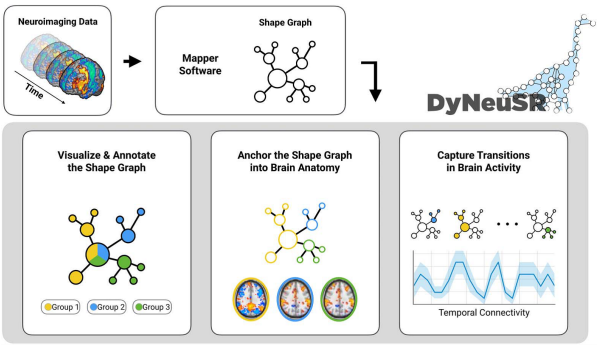

Figure 2.6.1: DyNeuSR overview. After the Mapper graph, DyNeuSR annotates metadata mapping the graph nodes into neurophysiology, hence capturing temporal variations
[4].

ure 2.6.2, we can observe spatial brain activation patterns obtained from clusters at relaxed state, 3 varied types of visual states and a controlled experiment cues (e.g. disorganized pictures).

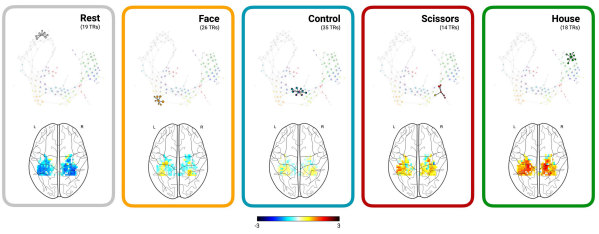

Figure 2.6.2: Mapping a shape graph into brain anatomy. Maps of brain activity from the ventral cortex approximated at varied time frames at different categories of visual stimuli from the Haxby study
[4].

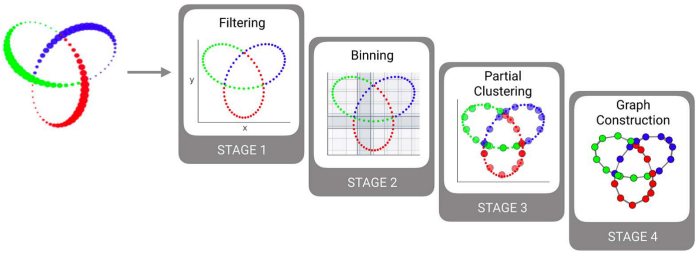

Figure 2.6.3: Visualization of the mapper stages[4]. This is Mapper visualization results obtained from synthetic data sampled from a 3-D trefoil knot

.

2.6.2 Mapper Algorithm

High-dimensional gets transformed into lower dimensional graphical forms when passed through the Mapper algorithm four stages. The graph shape shown in figure 2.6.3 was computed by the 1st and 2nd proportions of the three dimensional data. During the last stage, adjacent bins with common data points (i.e. with nonempty intersections) get connected together into a simple skeleton summary of the original dataset.

Compared to the standard dimensionality reduction techniques, the Mapper algorithm yields better results as it leverages the techniques from both clustering and dimensionality reduction of the initial highly featured space. As a demerit, standard DR techniques (DR) extrapolate TDPs to a minimal reduction state, from whence examination is done. Therefore, as a consequence, inasmuch as an accurate prediction is the motivation, this method always results to loss of some information from the dataset due to interpretability. Comparatively, Mapper takes advantage of the lens from the lower dimension to optimize the data in the initial highly-

dimensional space. The dataset is rich in illustration as it clearly reveals how standard dimensionality reduction techniques reveal very intricate patterns in two dimensions (as shown in diagrams 2.6.4 (A) and 2.6.4 (B). Find comparisons in diagram 2.6.4 C [4].

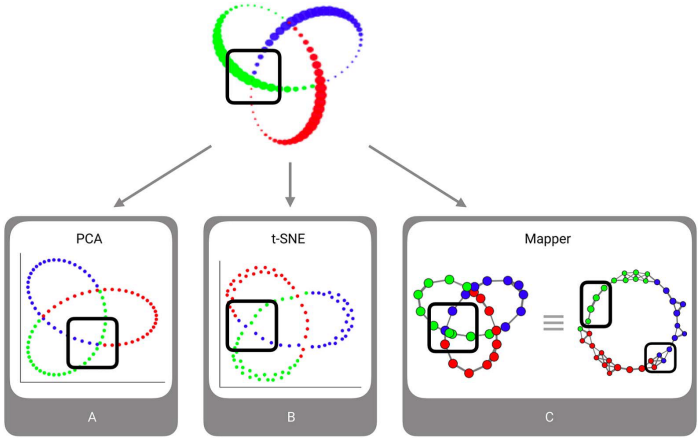

Figure 2.6.4: Visualization of the mapper advantages

. The figure gives a contrast description of synthetic 3-D trefoil knot revealed using traditional dimensionality reduction techniques such as linear (i.e., PCA) and nonlinear (i.e., t-SNE) with one generated using Mapper. Data points of a subset of a BTS was mapped from a higher dimensional space (top) to each of the lower dimension renditions (bottom). Notably, blue and green are separated by the third dimension in the high-dimensional space. PCA (A) and t-SNE (B) are both unable to resolve this separation-the blue and green points still remain in same position in the reduced dimension space. Conversely, Mapper (C) reveals these points as two separately disconnected clusters in the shape graph[4].

The database contains many triangulated meshed 3D shapes which were processed as follows. Letting P be the TDS out of which 4000 land-

mark are sampled through Euclidean maxmin procedure as illustrated in [12]. We denote this sampled set of points as $\mathbf{Y} = \{P_i, i \in L\}$, where L refers to the set of indices within the $\mathbf{Y}$. In order to expose this point cloud as an input to Mapper, the distance between the points in $\mathbf{Y}$ is computed as shown. To begin with, the adjacency metrix A for the set P is composed using the mesh information i.e. suppose P_i and P_j are connected on the given mesh, then $A(i,j) = d(P_i, P_j)$, such that $d(x,y)$ comprises to the distance between $x, y \in P$. The filter function $E_1(x)$ (setting $p = 1$) is chosen for use in applying the Mapper algorithm to this set of shapes as show in figure 2.6.5.

Eccentricity refers to the family of functions carrying defining the geometry of a TDS. Hence, having p together with $1 \leq p < +\infty$, we have $E_p(x) = (\frac{\sum_{y \in X} d(x,y)^p}{N})^{\frac{1}{p}}$ where $x, y \in \mathbf{X}$. The definition may be extended to $p = +\infty$ by setting $E_\infty = max_{x' \in X} d(x, x')$. Meanwhile, to minimize the bias effect as a result of the distribution of local features, a global clustering threshold is applied within all intervals. Our final result from this Mapper algorithm, is a graph i.e. a single filter function. To generate a visualization of this graph, GraphViz is employed and hence the results of a sampled shapes from the database are displayed in figure 2.6.6 with the following keypoints worth paying attention to [17].

(i). The Mapper algorithm has successfully recovered the graph symbolizing the skeleton shape with reasonable accuracy. Taking horse shape for instance, the three branches at the bottom, in both cases of the recovered graphs, signify the front two legs and the neck. Besides, the torso is symbolized by the blue colored section while the legs and the tail are represented by the top three branches of

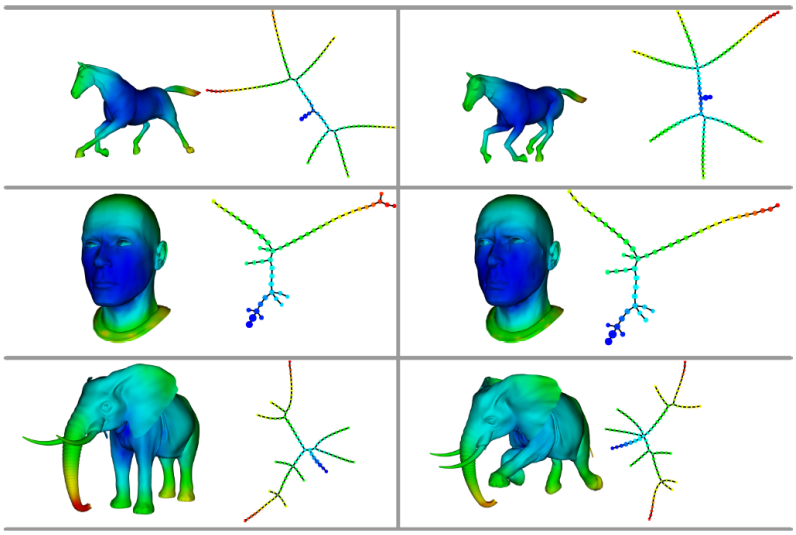

Figure 2.6.5: Every row displays two poses of the same shape together with the Mapper result. 15 intervals within the range of the filter function with 50% overlap is used to compute each Mapper instance
[17].

the graph. Ultimately, the skeleton would be recovered by bestowing the recovered graph with the mean position of the graph of the clusters they embody.

(ii). Different poses of similar shape have reliably identical Mapper results even though different shapes produce significantly dissimilar results. This is a revelation that this procedure retains certain intrinsic shape information, which is invariant to pose.

The above revelations unveiled by the Mapper algorithm under E_1 filter proposes that it may be a very instrumental tool for simplifying shapes, conducting database query and shape comparison tasks.

2.7 Research Gap Summary

Describing topological points is very intricate due to the nature of Big Data. This makes it difficult to locate BDS particularly in a general topological space setting. Because of the $T_2 - axiom$ in T_2-space, it is even more difficult to locate these BDS in Hausdorff spaces. In spite of the remarkable efforts put up by traditional techniques in data analysis, they have not always kept up with the exploding data quantity and complexity since they often depend on exceedingly simplistic assumptions and approximations during computations. Besides, they do not pay attention to the arbitrariness within this TDSs as well as the underlying instability within the topological datasets. Consequently, most of these techniques are exploratory, lacking the efficiency to distinguish what is sometimes called the "topological noise" from information of interest. The Vietoris-Rips complex for a parameter t has been so ubiquitously used to build a useful simplicial complex to mirror the data structure and utilizes the original data as the vertex set. The bone of contention, however, has always been how to choose the t parameter such that the Rips complex reveals the structure of the underlying data set. It is precisely this question that appeals to our conscience of thought towards persistence diagram as a topological signature of the dataset. Two metrics have commonly been used to measure the similarity of those objects: the bottleneck and Wasserstein distances. Each works by matching points of 1 diagram with points of another diagram while allowing the match to be done with the diagonal if necessary. The Mapper algorithm does have some limitations however. The topology of shape graphs is to a large extent dependent on whether filter function chosen is linear or nonlinear, as well as the

resolution or gain parameters. Also, comparison of the Mapper graphs computed from different filter functions is still very obscure. As much as topological models have been suggested, not many of them can be compared directly with the irregular structures derived from the big data sets. Consequently, the incapacitation in making these direct comparisons has proven to be a potential impediment to statistical validation of the Mapper results. Finally, previous studies done on TDA have had very little focus on the application of Hausdorff spaces.

Chapter 3

RESEARCH METHODOLOGY

3.1 Introduction

We, in this chapter, present the techniques, materials, and tools that we used to achieve our objectives. These included; Big Data Sets, Separation criterion of Hausdorff Spaces, AI and ML techniques, development of algorithms and simulations using python programming language. BD is described as hugely large, highly composite datasets to be analyzed by traditional techniques, but might reveal structure, patterns, relationships, shapes, when computational analysis methods are involved.

3.2 Separation criterion of Hausdorff Spaces

A Hausdorff space ($\mathbf{T}_2 - \mathbf{Space}$) refers to a topological space with which disjoint neighborhoods are possessed by any two distinct points of $\mathbf{X}$. A

space is qualified as Hausdorff when any of its two distinct points can be "housed off" independently into two open neighborhoods. Despite the development of fresh techniques to gather, store, and analyze large quantities and highly dimensional BDS, shape identification and interpretation has increasingly proven extremely challenging to visualize. Since more structure exists to be mined from this BDS, than can be revealed by the traditional methods, an extremely remarkable new method of shape identification in these BDS is the application Hausdorff spaces in Topological Data Analysis (TDA).

The Hausdorff spaces normally have the following two underlying features.

(i). Any two points can be separated, i.e. to an extent, they are far apart.

(ii). Every point can be approached very closely from other points, such that suppose we take a sequence of points, then they will normally gather around a point.

Many topologists believe that it is not all worth studying topological spaces that are not Hausdorff. Hausdorffness relates to the idea that picking any pair of distinct points, you must expand the points to disjoint open subsets. The fundamental idea is that: the notion of being close to the point is represented by disjoint open sets. Besides, nothing can come quite closer to two points when they are separated by disjoint open sets.

Theorem 3.1. ***Subspace of Hausdorff Space is Hausdorff***
Let $\mathbf{T} = (\mathrm{S}, \tau)$ *be a topological space which is* T_2. *Let* $\mathbf{T}_H = (\mathrm{H}, \tau_H)$,

where $\emptyset \subset \mathrm{H} \subseteq \mathrm{S}$, *be a subspace of* T*[31]. Then* $\mathbf{T}_H$ *is a* T_2 *(Hausdorff) space. This is, the property of being a* T_2 *(Hausdorff) space is hereditary.*

Hausdorff spaces will be very instrumental in the characterization of data points and locating of the Big Data Sets. This includes the development of the theorems, Lemmas and propositions, given the Big Data Sets. To generate more meaning from the Big Data Sets, we need to apply Artificial Intelligence and Machine Learning techniques as described in the next section.

3.3 AI and ML Techniques

Artificial Intelligence (AI) refers to an extensive branch of computer science involved in the theory and development of smart computer systems having the capability to perform tasks that usually require human intelligence. e.g. natural languages translation, voice recognition and visual perception. ML refers to a discipline of AI that utilizes computer algorithms to learn, experience, adapt and automatically improve without human programming. One such techniques of Machine Learning used to analyze and feature engineer the Big Data Sets is the t-SNE unsupervised Machine Learning algorithm. The next section details the t-SNE machine learning algorithm.

3.3.1 The t-SNE algorithm

t-SNE refers to feature engineering algorithm to unsupervised machine learning algorithms (Applied AI) like k-means. t-SNE can be described as a non-linear DR algorithm instrumental in highly dimensional data exploration. Dimensionality reduction refers to a linear or non-linear technique involved in mapping higher dimensional to a low-dimensional space while preserving local features within the primary space. Reducing the dimensions to a lower dimension by achieves the preservation of two things: If data points are close by in the high dimensional space, it tries to retain that closeness in small dimension space. If points are far apart, it also tries to keep them a part in a smaller dimensional space. So it preserves closeness and farness in the space. And it does that by applying attractive forces between points that are close and repulsive forces to points that are apart. These forces are applied repeatedly to all the points in the space for a number of iterations between points closer and those far apart. T-SNE is therefore important in showing clusters of data that are similar and close. We choose t-SNE because linear dimensionality reduction algorithms like PCA, emphasize on positioning contrasting TDPs distantly separated in a lower dimensional space. However, achieving representation of highly dimensional TDSs on a low dimension, indistinguishable datapoints must be mapped closer together and this can only be achieved by non-linear algorithms unlike their linear counterparts. t-SNE is an enhancement over the (SNE) algorithm[30]. To implement the t-SNE ML algorithm, the python programming language has proved very instrumental in Machine Learning, given its vast libraries and frameworks for advanced computing and visualization. The t-SNE Machine Learning

algorithm will be very applicable in establishing the distribution patterns of the topological data points within a Hausdorff space. The next section describes the python libraries, distributions and the computing requirements applicable in Machine Learning for our study.

3.4 Algorithms and Simulations in Python

Python is popular programming language that is widely used in Artificial Intelligence communities. Python possesses simplicity and readability features for its syntax and this reduces the time taken to test complicated algorithms through minimal code in comparison to the existing languages[23]. Python prides itself with a great numbers of rich library modules for Machine Learning. Furthermore, python commands overwhelming community of developers globally, who generously share troubleshooting and debugging tips through online platforms. The ubiquitous nature of Python has granted its usage in nearly all research institutions as well as commercial applications of Deep Learning and Machine Learning. Jupyter Notebook is a computing notebook environment operated on a interactive web browser. It is a component of the Anaconda Navigator. Some of the python libraries include; pandas, sklearn, matplotlib, seaborn, plotly, cufflinks, TSNE, KMeans. Some of these libraries are resource intensive especially the ones that implement Machine Learning algorithm to run through over thousands of rows of data for hundreds of iterations, in just a few minutes. Therefore, the laptop specifications to be used to run these python algorithms must possess a powerful processor of Core i7, 16GB of RAM memory, over 2.3 GHz speed, and a faster

500GB Solid State Drive storage. The next section presents the Big Data Sets that will be used in our study.

3.5 Big Data Sets

The datasets used in this study are downloaded from Kaggle.com, an open source community of data scientists, ML, and a huge published Repository of BD sets. We specifically focus on a collection of time series COVID-19 datasets of cases reported from all countries of all the six world continents (Africa, Europe, Asia, North America, South America and Oceania) starting February 24^{th}, 2020 until June 28^{th}, 2021. The dataset contain 98,904 rows and 60 columns. The huge volume of this dataset qualifies it as a candidate of a Bid Data set. We denote the COVID-19 data set as $\mathbb{X}$ throughout the study. Besides, python also has the ability to randomly generate larger data sets in a simulated environment for analysis purposes. The next section takes us through the initial cleaning of the raw data and exploratory data analysis, in order to get maximum output from our data set.

3.5.1 Exploratory Data Analysis and Data Cleaning

The raw data has to go through several processes before the final analysis. The data is first imported into the python using the Pandas library. We then expose the data through the process of shaping to determine the initial high dimension, we check the data distribution, wrangling and initial visualization processes. Subsequently, Exploratory Data Analysis

(EDA) follows, where we perform structural modification, further analysis and visualization on the data using line graphs and surface plots using various python libraries. After EDA, we perform a very crucial stage on the data known as feature engineering. This involves feature selection, dimensionality reduction using techniques like backfilling, forward-filling, data encoding, best column formation, variance and means computations. These processes achieve data cleaning which assists us to manage handling of outliers and excessive null values; which TDA refers to as "topological noise". At this point, the data can finally be exposed to the ML Algorithms; either as Unsupervised, Supervised, Semi-supervised, or reinforcement ML algorithms. During our study, we will however restrict ourselves to t-SNE which is a feature engineering algorithm to unsupervised machine learning algorithm as a branch of Applied Artificial Intelligence. In the next section, we embark on t-SNE as a Machine Learning algorithm.

In the next chapter, we detail the results of our study as a result of the methodology described in this chapter.

Chapter 4

RESULTS AND DISCUSSION

4.1 Introduction

In this chapter, we give the results of our study. For the first objective, we characterize topological data points in Hausdorff spaces. For the second objective, we locate Big Data sets in Hausdorff spaces. For the third objective, we establish distribution patterns of topological data points in Hausdorff spaces. This results utilizes Covid-19 Big Data Sets, Python libraries, Artificial Intelligence and Machine Learning techniques, and Hausdorff spaces which provide higher efficiency, more reliability, and with a more faster algorithmic complexity.

4.2 Topological Data Points

In this section, we characterize TDPs in a Hausdorff space. From this point, we consider $\mathbf{X}$ as a Hausdorff space and $\mathbf{H}$ is a subspace of $\mathbf{X}$, unless otherwise stated. We begin with the following proposition.

Proposition 4.1. *Let* $\mathbf{H}$ *be a TDP space and* $\mathrm{a} \in \mathbf{H}$ *such that* $\mathbf{H} \setminus \{\mathrm{a}\} = \mathbf{P}\dagger\mathbf{Q}$. *Then* $\{\mathrm{a}\}$ *is open or closed. If* $\{\mathrm{a}\}$ *is open then* $\mathbf{P}$ *and* $\mathbf{Q}$ *are closed in* $\mathbf{H}$ *and if* $\{\mathrm{a}\}$ *is closed, then* $\mathbf{P}$ *and* $\mathbf{Q}$ *are open.*

Proof. Since $\mathbf{P}$ is both open and closed in $\mathbf{H} \setminus \{a\}$, we have an open subset $\mathbf{R}$ of $\mathbf{H}$ such that $\mathbf{P} = \mathbf{R} \cap (\mathbf{H} \setminus \{a\}) = \mathbf{R} \setminus \{a\}$ and there exists a closed $\mathbf{Z}$ of $\mathbf{H}$ in which $\mathbf{P} = \mathbf{Z} \cap (\mathbf{H} \setminus \{a\}) = \mathbf{Z} \setminus \{a\}$. Therefore, $\mathbf{P} = \mathbf{R} \setminus \{a\} = \mathbf{Z} \setminus \{a\}$. Analogously, $\mathbf{Q} = \mathbf{R} \setminus \{a\} = \mathbf{Z} \setminus \{a\}$. □

This proposition leads to the characterization of topological data points in terms of compactness as seen in the next level.

Lemma 4.2. *Let* $\mathbf{H}$ *be a TDP space and* $\mathrm{a} \in \mathbf{H}$. *If* $\mathbf{H} \setminus \{\mathrm{a}\} = \mathbf{P}\dagger\mathbf{Q}$ *then* $\mathbf{P} \cup \{\mathrm{a}\}$ *is compact.*

Proof. Without loss of generality, let $\mathbf{W}$ and $\mathbf{V}$ be connected and compact subsets of $\mathbf{H}$ in which $\mathbf{P} \cup \{a\} = \mathbf{W}\dagger\mathbf{V}$. Let $a \in \mathbf{W}$. Then $\mathbf{V} \subseteq \mathbf{P}$. Now $(\overline{\mathbf{Q} \cup \mathbf{W}}) \cap \mathbf{V} = (\overline{\mathbf{Q}} \cap \mathbf{V}) \cup (\overline{\mathbf{W}} \cap \mathbf{V}) = \emptyset$. So $(\overline{\mathbf{Q} \cup \mathbf{W}}) \cap \mathbf{V} = \emptyset$ and consequently, $(\mathbf{Q} \cup \mathbf{W}) \cap \overline{\mathbf{V}} = \emptyset$ implying $\mathbf{H} = (\mathbf{Q} \cup \mathbf{W})\dagger\mathbf{V}$. □

In the next level, we give our main theorem for the first objective followed by its proof.

Theorem 4.3. *Let* $\mathbf{H}$ *be a TDP space and* $\mathrm{a} \in \mathbf{H}$. *If* $\mathbf{H} \setminus \{\mathrm{a}\} = \mathbf{P}\dagger\mathbf{Q}$ *and if every point of* $\mathbf{P}$ *is TDP in* $\mathbf{H}$ *then* $\mathbf{P}$ *has at least one closed point.*

Proof. Suppose that $\mathbf{P}$ is compact then by Proposition 4.1, $\mathbf{P} \cup \{a\}$ is compact. So $\{a\}$ is closed. Let $z_0 \in \mathbf{P}$. By Lemma 4.2, $\mathbf{H} \setminus \{z_0\} = \bigcup_{z=\mathbf{P}, z \neq z_0} \{\{a, z\} \cup (\mathbf{Q} \cup \{a\})\}$ is also compact. This contradicts the earlier hypothesis that a is TDP point of $\mathbf{H}$. □

Remark 4.4. All TDP spaces are connected spaces. However, a finite topological space is not a TDP.

4.3 Location of Big Data Sets

In this section, we locate Topological Data points of Big Data Sets in a Hausdorff space. We state the following proposition.

Proposition 4.5. *Let* $\mathbf{H}$ *be a TDP space. The set* $\mathbf{A}_0$ *of all condensation points of* $\mathbf{H}$ *is a TDS which is infinite.*

Proof. Let a_1, a_2,... be a sequence of distinct condensation points in $\mathbf{H}$. By induction, we have a condensation point a_0 in $\mathbf{A}_0 \subseteq \mathbf{H}$. But a_0 is a TDP of $\mathbf{H}$. So we have open TDS $\mathbf{W}_1$ and $\mathbf{V}_1$ of $\mathbf{H}$ such that $\mathbf{H} \setminus \{a_1\} = \mathbf{W}_1\dagger\mathbf{V}_1$. Suppose that $a_1, a_2, ..., a_n$ are in $\mathbf{H}$ and open subsets $\mathbf{W}_i$ and $\mathbf{V}_i (i \in \mathbf{N})$ are picked such that $\mathbf{H} \setminus \{a_i\} = \mathbf{W}_i\dagger\mathbf{V}_i$, where $i = 1, ..., n$. Clearly, by induction and considering $\mathbf{W}_{i+1}$ and $\mathbf{V}_{i+1}$, the set $\mathbf{A}_0$ of all condensation points of $\mathbf{H}$ is infinite. □

Proposition 4.5 takes us to characterization of the size of the sets. We give the size of the data set in the next result as follows.

Lemma 4.6. *Let* $\mathbf{H}$ *be a TDP space. Then* $Card\mathbf{H} = \infty$.

Proof. By Hausdorff Maximal Principle (HMP) and by Proposition 4.5, the proof is complete. □

To further demonstrate the characterization of the topological Data Points within a Hausdorff space, we conclude with the following theorem.

Theorem 4.7. *All TDS in a TDP space are arbitrarily distributed if they are* $\boldsymbol{T}_2$. *Moreover, each TDS has at least two TDPs with closed TDS which are singletons.*

Proof. Let $\mathbf{H}$ be a TDP space and let $\mathbf{H}_1$ and $\mathbf{H}_2$ be TDP subspaces of $\mathbf{H}$. Then it implies that if $\mathbf{H}_1$ and $\mathbf{H}_2$ are both empty then trivially we are done. Let $\mathbf{H}_1$ and $\mathbf{H}_2$ be non-empty. It remains to show that $\mathbf{H}_1 \cap \mathbf{H}_2 = \emptyset$ and hence it is $\mathbf{T}_2$. To see this, consider $a_1 \in \mathbf{H}_1$ and $a_2 \in \mathbf{H}_2$ such that $a_2 \notin \mathbf{H}_1$ and $a_1 \notin \mathbf{H}_2$. Clearly, $\mathbf{H}_1 \cap \mathbf{H}_2 = \emptyset$, hence it is Hausdorff. Now we show that $\mathbf{H}$ has at least two TDPs. Let $\mathbf{H}$ be such that it has at most one TDP. Let $a_1 \in \mathbf{H}$ and $\mathbf{H} \setminus \{a_1\} = \mathbf{P}_0 \dagger \mathbf{Q}_0$ for some $\mathbf{P}_0, \mathbf{Q}_0$, which are subsets of $\mathbf{H}$. Since $\mathbf{H}$ has only one TDP then either $\mathbf{P}_0$ or $\mathbf{Q}_0$ has TDPs. By proposition 4.5, $\mathbf{P}_0$ has some condensation point of $\mathbf{H}$ say a. Let $\mathbf{H} \setminus \{a\} = \mathbf{P} \dagger \mathbf{Q}$. Without loss of generality, let $a_0 \in \mathbf{Q}$. By Hausdorff Maximal Principle, there is an optimal chain $\mathbf{C}$ in $\mathbf{S}$ of $\mathbf{H}$ such that for some $\mathbf{U}_\alpha$ of $\mathbf{S}$, $\bigcup_{\alpha \in \Lambda} \mathbf{U}_\alpha \in \mathbf{H}$. Hence, $\mathbf{H}$ is compact. Let $\mathbf{V} = \bigcup_{\alpha \in \Lambda} \mathbf{U}_\alpha$, then by Lemma 4.6, we can get at least two points of $\mathbf{H}$ which give a subcover for $\mathbf{H}$. Since the subcovers are open by Heine-Borel Property (HBP), each set forms a singleton set. □

In the next Section, we demonstrate the distribution patterns of TDPs using real life scenarios of Big Data Sets. Each TDP represents a Covid-19 case and the TDSs represents subsets of the Hausdorff Space **X**.

4.4 Distribution Patterns of TDPs

4.4.1 Data Set Repository

The datasets used in this study are downloaded from Kaggle.com, an open source community of data scientists, ML, and a huge published Repository of BD sets. In our coding, the world is represented by a Hausdorff space **X** and subspaces of **H** represent regions or countries, unless otherwise stated. These subspaces represent Big Data sets. We specifically focus on a collection of time series COVID-19 datasets of cases reported from all countries of all the six world continents (Africa, Europe, Asia, North America, South America and Oceania) starting February 24^{th}, 2020 until June 28^{th}, 2021. The dataset contained 98,904 rows and 60 columns. The huge volume of this dataset qualifies it as a candidate of a Bid Data set. The next section describes the connectedness of our data set. The data can be downloaded from:
https://www.kaggle.com/imdevskp/corona-virus-report?select=country$_w ise_l atest.csv$
$https://www.kaggle.com/kalilurrahman/covid19-coronavirus-dataset-by-owid?select=owid-covid-data.csv$

4.4.2 Connectedness of the data points

Given $\mathbf{H}_1$,$\mathbf{H}_2 \subseteq \mathbf{H}$, such that $\mathbf{H}_1 \cap \mathbf{H}_2 = \emptyset$, then collections generated by $\mathbf{H}_1$ and $\mathbf{H}_2$ might be equated to non-empty intersections, and these Hausdorff properties come in handy in building a SC. As a result, this result leads to a multiresolution" map of the TDS. Connectedness of points in a topological space means nonexistence of separation between the points on the topological space. Let $\mathbf{X}$ be our COVID-19 data set throughout this study. We applying a non-linear fit on $\mathbf{X}$ by reducing the dimensionality while the geometric structure, shape and connectivity of our data set remains preserved.

4.4.3 2D Line Graph Visualizations

After Exploratory Data Analysis (EDA) and Dimensionality reduction on our COVID-19 data set, denoted by $\mathbf{H}$, we produce a few line graph visualizations to initially reveal the relationship between the total confirmed, Recovered and Deaths cases across the six world continents. A visualization of the Total Confirmed COVID-19 cases in the world is revealed at a glance through the line graph in figure 4.4.1. Figure 4.4.2 gives a line graph visualization of the total Recovered cases in the world. The total number of Deaths is visualized by the line graph in figure 4.4.3. Finally, figure 4.4.4 reveals a combined line graph visualization of both the Total confirmed, Recovered and Death cases of COVID-19 in the world.

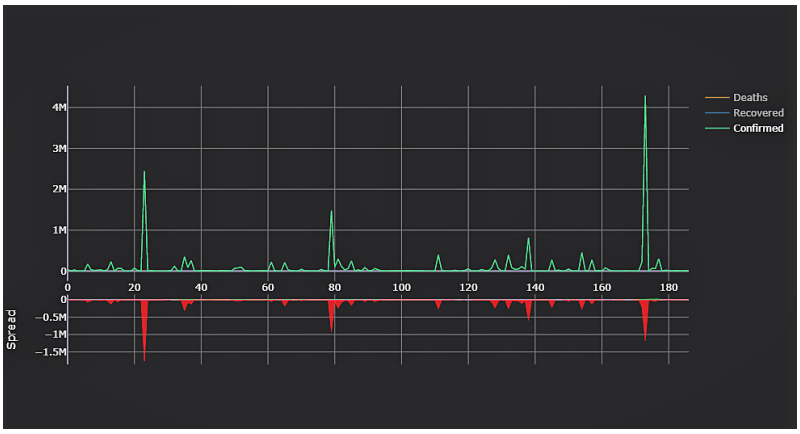

Figure 4.4.1: Line plot visualization of dataset **H** revealing the total confirmed Covid-19 infection cases globally.

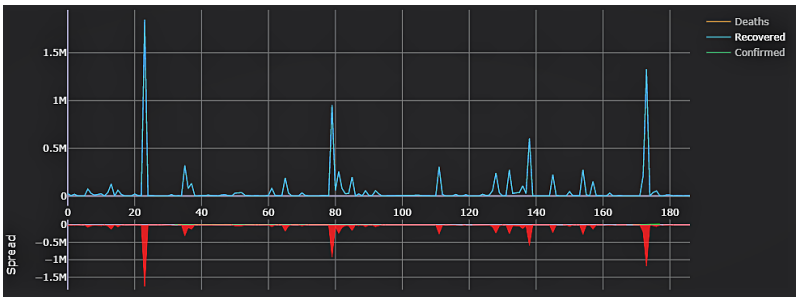

Figure 4.4.2: Line plot visualization of dataset **H** revealing the total recovered cases globally.

4.4.4 3D Surface Plot Visualizations

In addition, after performing Exploratory Data Analysis and Dimensionality reduction on our COVID-19 data set denoted by **H**, we compute surface plot visualizations to reveal topological structure and graphical properties from our data set **H**. We use python libraries to produce surface plot visualizations which revealed the structured shape between the total confirmed, total Recovered and total Deaths cases across the

Figure 4.4.3: Line plot visualization of dataset **H** revealing the total confirmed death cases globally.

six world continents. Figure 4.4.5 displays elevated numbers of confirmed cases globally, slightly lower deaths per hundred cases, and relatively fewer New Cases during the period when the data was collected. Figure 4.4.6 visualizes a surface plot of the Covid-19 dataset displaying higher numbers of New cases globally, slightly moderate New deaths cases, and relatively fewer Death Cases globally, during the period of data collection. Figure 4.4.7 displays a surface plot visualization of the Covid-19 dataset displaying extremely higher numbers of New Deaths globally, moderately higher New Recovery cases, and relatively lower Deaths per hundred Cases globally. Each TDP is enclosed around a ball at the center at a point with a radius ε. While ε increases in size, the cloud gradually ceases to appear as secluded points, but slowly gains shape. As the output increases in size, a single shapeless blob emerges. This events eventually generates a SC, which starts with vertices as TDPs. An edge is inserted between the two balls as they intersect. Meanwhile, a face bounded by the three edges is added as 3 balls intersect. With the onset of these events, highly

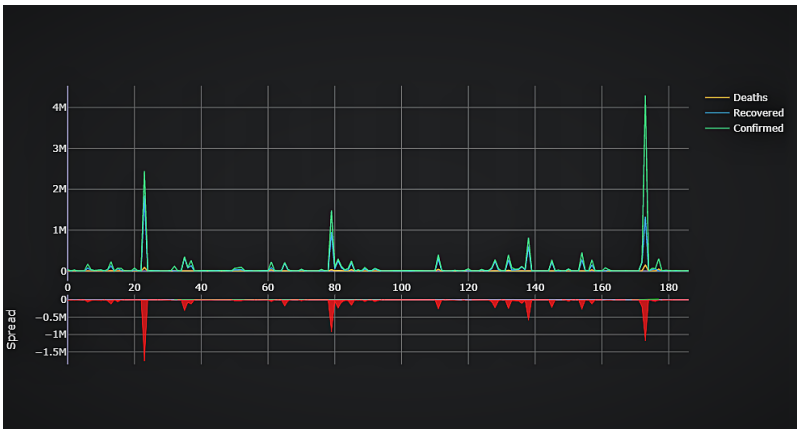

Figure 4.4.4: A combined Line plot visualization of dataset **H** revealing the total confirmed, recovered and death cases globally.

dimensional n-faces with n+1 intersecting emerges. Taking **H** and draw a ball $\mathbf{H}(a_1, r)$ around each point a_1 for some small initial radius r (figure 4.4.8). As r increases at intervals, we can establish how **H** is connects at different radii by observing different angle of snapshots of the 3-D outputs. We obtain a 3-dimensional scatter plot non-linear random shape of the data at an initial default radius of 12 as shown in figure 4.4.8. At this point, we will keenly observe whenever the initial intersection of the closures of $\mathbf{H}(a_1, r)$ and $\mathbf{H}(a_2, r)$ on a point within the plane. Without loss of generality, we can confidently state that a_1 and a_2 are $2r$ distanced apart. Connecting them with an edge, we get a visual graph. We obtain a 3-dimensional scatter plot non-linear random shape of the data at radius of 15 as shown in figure 4.4.9. Initially, we see observe sparse clusters of edges connecting vertices. As r increases, more edges emerge in small connected clusters for a radius of 15. We obtain a 3-dimensional scatter plot non-linear random shape of the data at radius of 15 as shown in figure 4.4.10. We now begin to see the first instance of structure on

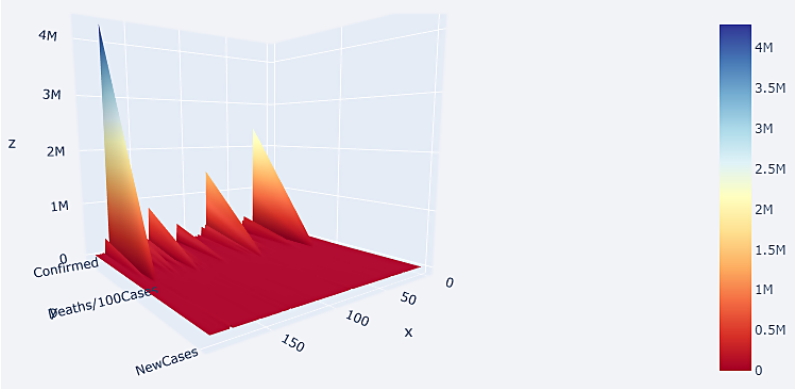

Figure 4.4.5: A 3D surface plot visualization of the Covid-19 dataset displaying higher numbers of confirmed cases globally, slightly lower deaths per hundred cases, and relatively fewer New Cases globally.

the data which can we can call the connectivity information of the data. These clusters reveal valuable information about the underlying features such that when some initial conditions of the experiment are fulfilled, the data points are more likely to assemble in a certain manner. We obtain a 3-dimensional scatter plot non-linear random shape of the data at radius of 30 as shown in figure 4.4.11. Thereafter, notable clusters may begin to emerge, varying from those achieved with smaller radii, basically encompassing them. As the radius of the balls increase, a single component of **H** is finally observed for the very first time. The component can among other things be a loop, a chain, have holes or loops, have multiple flares, or another structure. We obtain a 3-dimensional scatter plot non-linear random shape of the data at radius of 40 as shown in figure 4.4.12. With the further increase of the radius r, the loop in the data begins to emerge as seen in figure 4.4.13. So as the radius r increases towards r^*, the balls around the points a_1 and a_2 intersect and the final edge emerges to com-

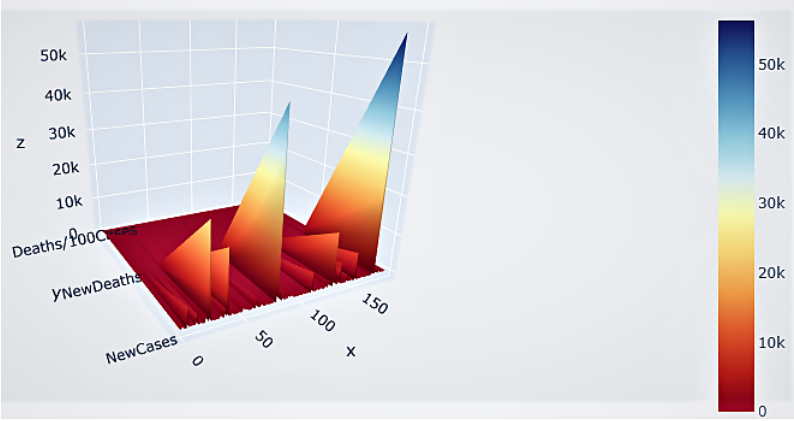

Figure 4.4.6: A 3D surface plot visualization of the Covid-19 dataset displaying higher numbers of New cases globally, slightly moderate New deaths cases, and relatively fewer Death Cases globally.

plete the loop. When eventually this radius is attained, further increment of the radius reveals no further structure for a significant time period. We therefore conclude that the ranges of the radius is $r > r^*$ as shown. We obtain a 3-dimensional scatter plot non-linear random shape of the data at radius of 50 as shown in figure 4.4.13.

4.4.5 t-SNE Clusters

Before we apply the t-SNE algorithm, we first compute a quick visualization of three categories within the global covid-19 dataset. These groups include; the cardiovascular death rate, diabetes prevalence and the population aged 70 years and older. The 3D surface plot on figure 4.4.14 visualizes a common relationship between the three groups. From the data set, we also compute a 3D visualization to reveal the shape of three more categories, i.e. total confirmed, total recovered and the the total

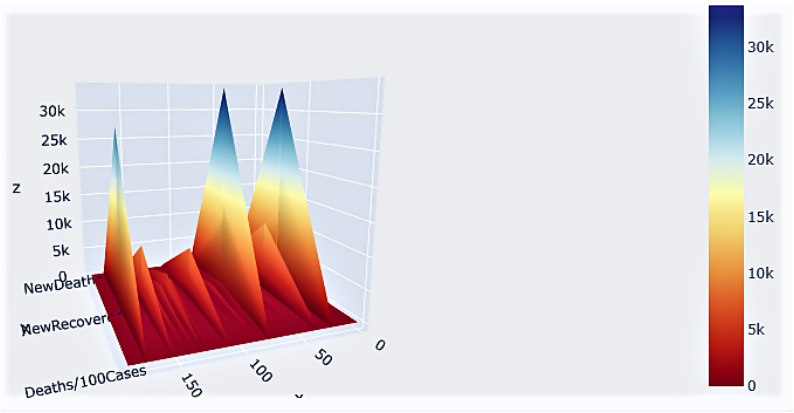

Figure 4.4.7: A 3D surface plot visualization of the Covid-19 dataset displaying extremely higher numbers of New Deaths globally, moderately higher New Recovery cases, and relatively lower Deaths per hundred Cases globally.

deaths among the six continents as shown in figure 4.4.15. After the Covid-19 dataset is subjected to the t-SNE algorithm, a feature engineering algorithm to unsupervised machine learning algorithms, we obtain the following three clusters distributions; at the global, Continental (Africa) and country (Kenya) levels. Within the clustered distributions, a shape is revealed outlining the distribution pattern of the data points within the Covid-19 data sets. Also included, are outliers (Topological "noise" (TN)). The outliers include extremely high or extremely low figures within the datasets. Also noted within the distribution spaces, condensed points and sparsely distributed data points. The global distribution pattern is revealed in figure 4.4.16.

You can find our code at the links below. However, to run the code to view the visualizations, you might need to install Anaconda Navigator, a python distribution library, for Machine Learning, install the relevant

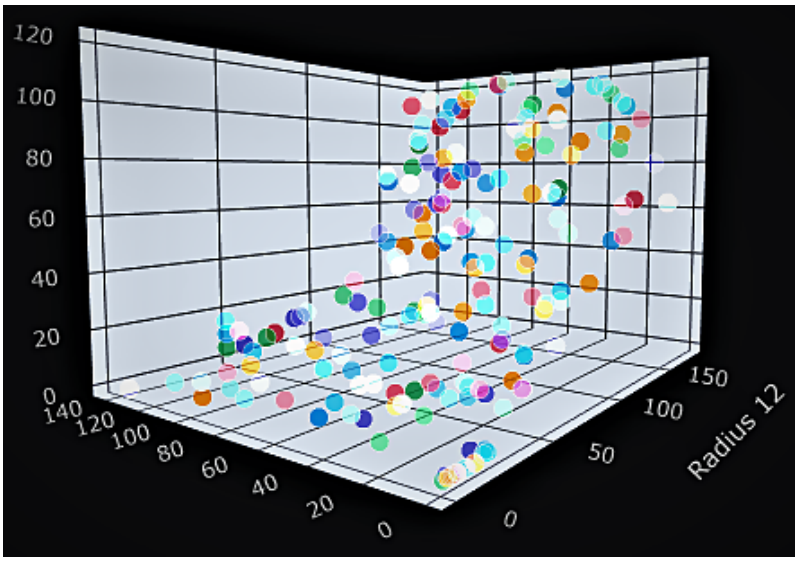

Figure 4.4.8: 3D scatter plot of the sample **H** of data points in at initial default radius of 12.

libraries, then download the CSV datasets from the kaggle.com repository:
http://localhost:8888/notebooks/Hausdorff%203D%20A.I%20Visualizations.ipynb
http://localhost:8888/notebooks/Hausdorff%20Machine%20Learning.ipynb

Figure 4.4.17 reveals the distribution pattern of the datapoints within Africa as a continent. Finally, figure 4.4.18 reveals the distribution pattern of the data points of the Kenyan portion of the global portion of the Covid-19 data set.

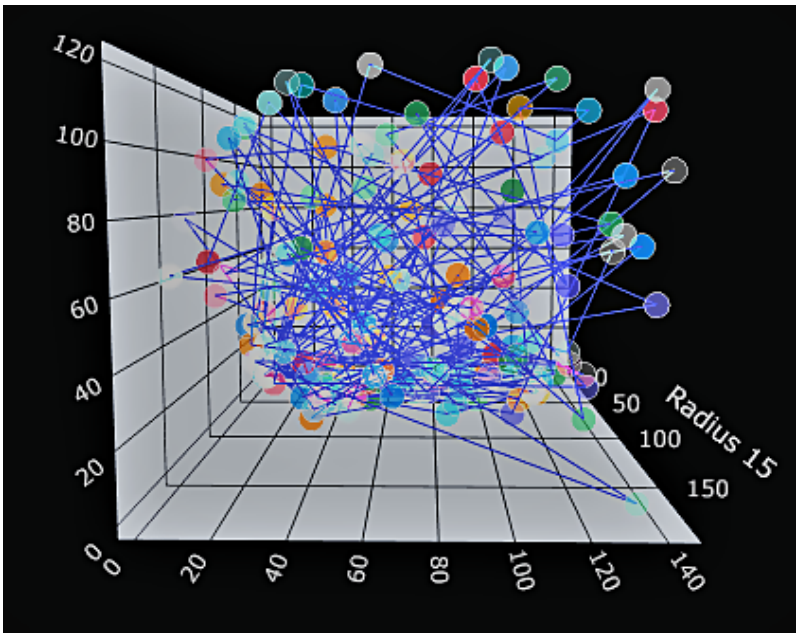

Figure 4.4.9: 3D scatter plot of the first radius increment to 15.

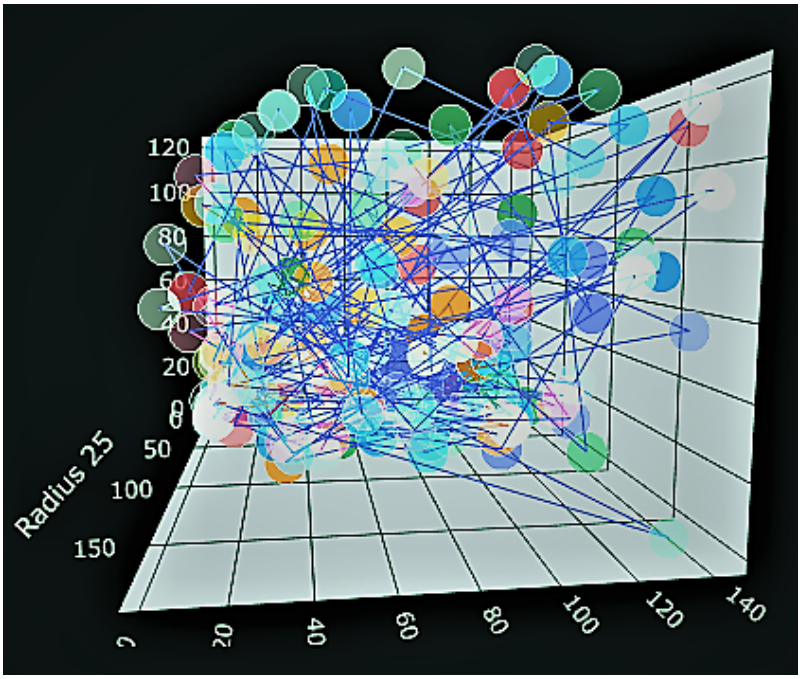

Figure 4.4.10: Denser 3D scatter plot clusters due to increasing connections at radius of 25.

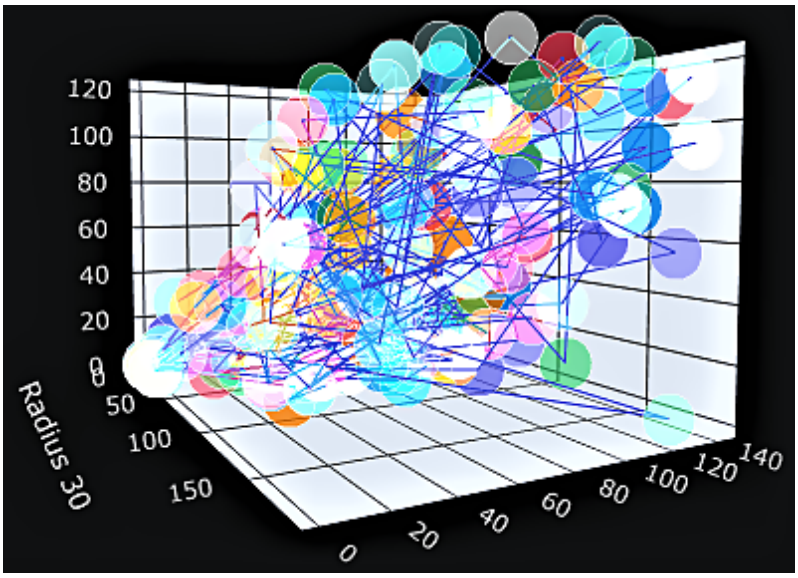

Figure 4.4.11: 3D scatter plot non-linear random shape of the data **H** at radius of 30.

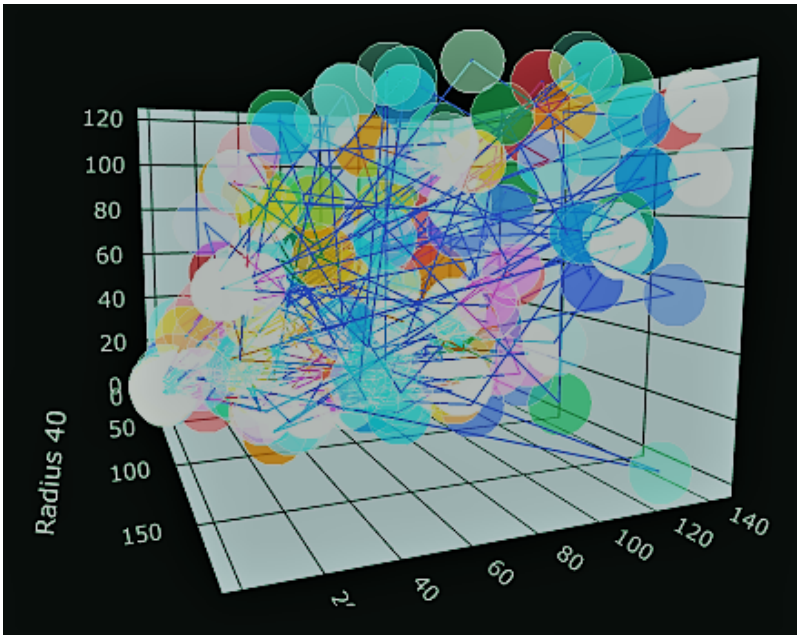

Figure 4.4.12: 3D scatter plot non-linear random shape of the data **H** at radius of 40

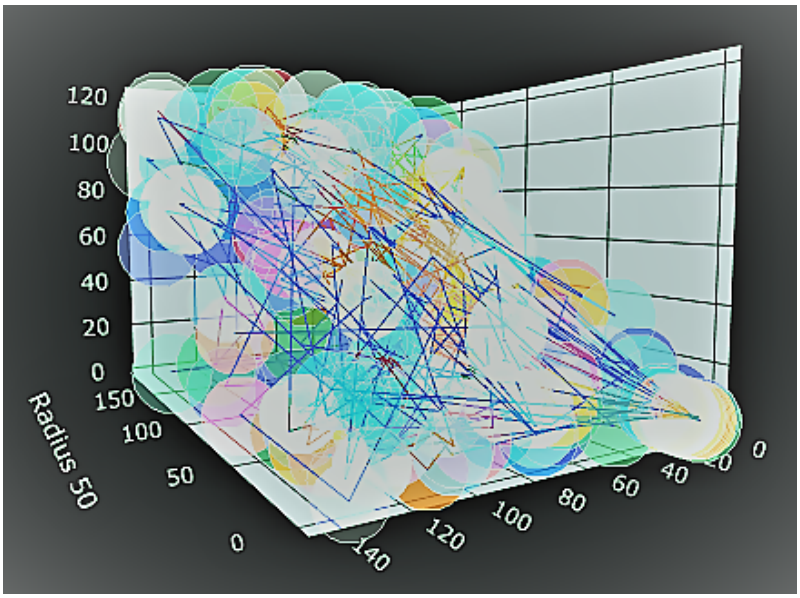

Figure 4.4.13: 3-D largest scatter plot non-linear random shape of the data **H** at radius of 50.

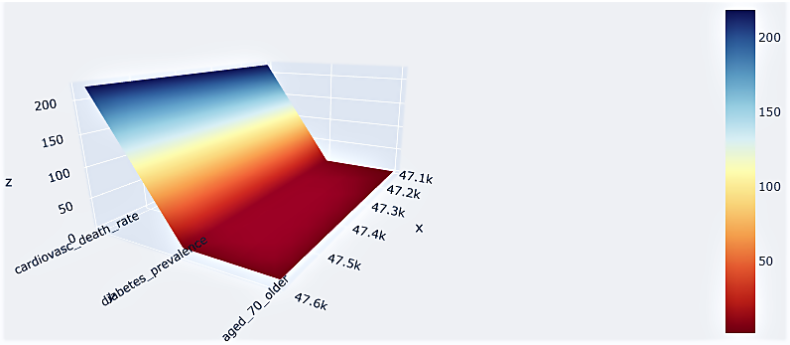

Figure 4.4.14: 3D Surface plot visualization revealing how cardiovascular death rate, diabetes prevalence and the population aged 70 years and older are related.

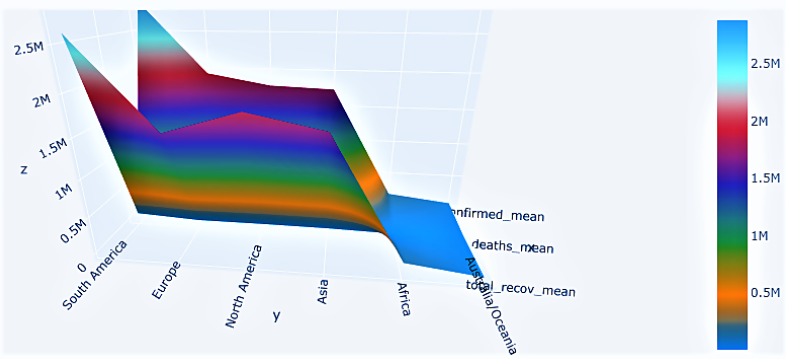

Figure 4.4.15: 3D Surface plot visualization revealing how the means of total confirmed, total recovered and the the total deaths among the six continents are related.

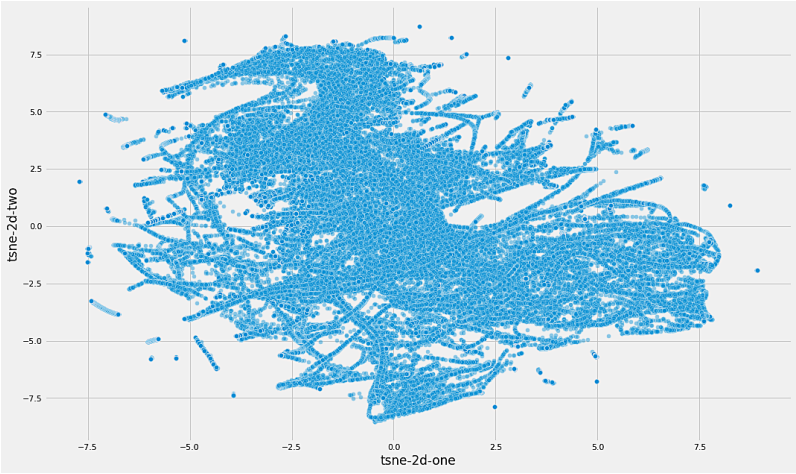

Figure 4.4.16: t-SNE generated distribution pattern of Covid-19 data points within statistics of the world.

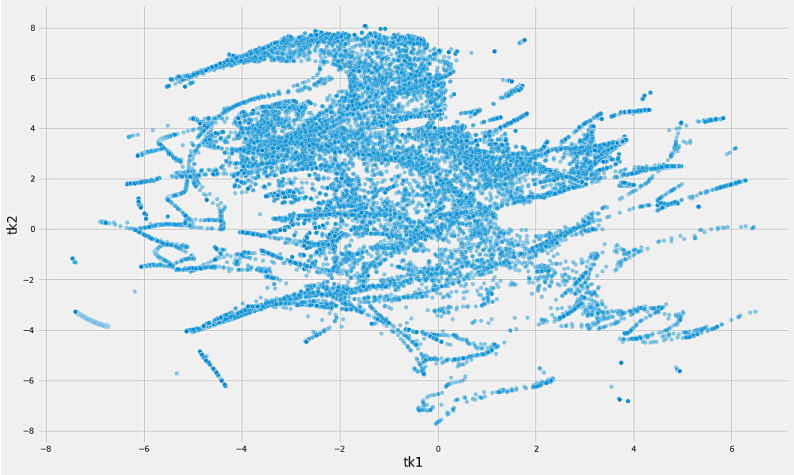

Figure 4.4.17: t-SNE generated distribution pattern of Covid-19 data points within statistics of the African continent.

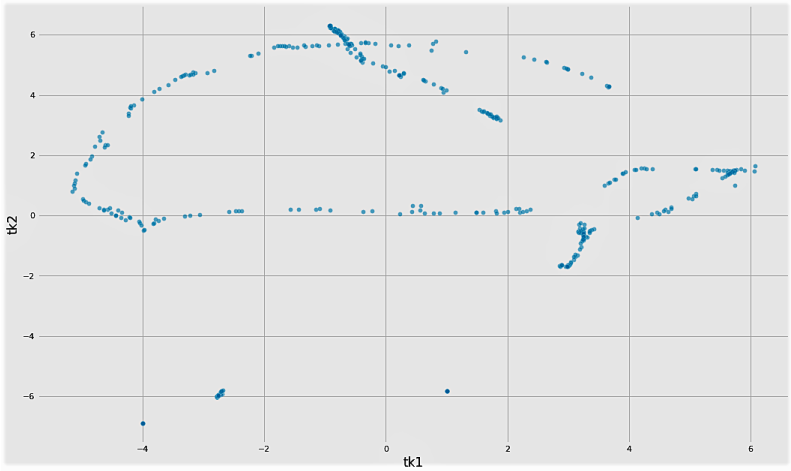

Figure 4.4.18: t-SNE generated distribution pattern of Covid-19 data points of Kenya.

Chapter 5

CONCLUSION AND RECOMMENDATIONS

5.1 Introduction

In this chapter, we give a conclusion and recommendations of our study as per the objectives in Section 1.4. We have considered characterization and location of topological data points in a Hausdorff space. Finally, we established the distribution patterns of topological data points using real life data sets of Covid-19.

5.2 Conclusion

In this section, we give a conclusion of our study. For the first objective, we have determined that topological data points form singleton sets which are closed. Besides, the set of topological data points is compact, and finally, the set of topological data points has at least one closed point.

As for our second objective, we have determined that the set of all condensation points of a topological data points is infinite. Also, the cardinality of a topological data set is infinite, and finally, topological data sets are arbitrarily distributed is Hausdorff spaces.

As for our last objective, we have managed to demonstrate the distribution patterns of topological data points within a Hausdorff space, using 3D visualizations and application of t-SNE Machine Learning algorithm clusters of the data set from all the six World continents, African continent and the country Kenya. From the distribution of the real life Covid-19 data set, the Coronavirus situation was densely distributed in Winter-prone regions like Europe, United States of America, and Canada.

5.3 Recommendations

From this study, we recommend that characterization of topological data points can be considered in Normal spaces.

We further recommend that location of Big Data Sets can be carried out in normal spaces.

Finally, we recommend that establishment of distribution patterns of topological data points can be done in normal spaces using data from other fields like health, business and social media.

References

[1] **Marr B.,** Big Data Using SMART Big Data Analytics and Metrics To Make Better Decisions and Improved Performance, *Library of Congress*, 2015.

[2] **Bremer P.,** A topological hierarchy for functions on triangulated surfaces, *IEEE Trans. Vis. Comp. Graph.*, 10(2004), 385-396.

[3] **Butler D.,** A world where everyone has a robot: why 2040 could blow your mind, *Nature 530*, (2016), 75-91.

[4] **Caleb G., Olaf S., Giovanni P., and Manish S.,** Generating dynamical neuroimaging spatiotemporal representations (DyNeuSR) using topological data analysis, *Network Neuroscience.*, 3(3)(2019), 763-778.

[5] **Carlsson G.,** Topological pattern recognition for point cloud data, *Acta Numerica*, 23(2014), 289-368.

[6] **Carlsson G.,** Topology and data, *Bull. Am. Math. Soc.*, 46(2009), 255-308.

[7] **Carrire M., and Oudot S.,** Structure and stability of the 1-dimensional mapper, *arXiv preprint*, 2015

[8] **Chen H., Chiang V., Storey**, Business intelligence and analytics: From big data to big impact, *MIS Q.*, 36(4)(2012), 1165-1188.

[9] **Colleen F.,** https://www.quora.com/What-does-Mike-West-think-about-Colleen-Farrellys-views-on-data-science (accessed on February 2021).

[10] **dataschools.education** *https://dataschools.education/about-data-literacy/what-is-data/*, (accessed on February 2021).

[11] **Dayten S.,** Introductory Topological Data Analysis. *Department of Mathematics and Statistics University of Victoria*, 2020.

[12] **De Silva V., Carlson G.,** Topological estimation using witness complexes, In Eurographics Symposium on Point-Based Graphics, *ETH, Zurich*, (2004), 157166.

[13] **Eselsbrunner, Letscher, Zomorodian,** Topological Persistence and Simplification, *Disc. Comp. Geom.*, 28(2002), 511-533.

[14] **Elizabeth M.,** A Users Guide to Topological Data Analysis, *Journal of Learning Analytics.*, 4(2), 47-61. http://dx.doi.org/10.18608/jla.2017.42.6

[15] **Frederic C., and Bertrand M.,** An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists, *National Institute for Research in Computer Science and Control*, 2017.

[16] **Frosini P.,** Measuring shapes by size functions. In Intelligent Robots and Computer Vision X: Algorithms and Techniques, *International Society for Optics and Photonics*, (1992), 1607.

[17] **Gurjeet S., Facundo M., and Carlsson G.,** Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition, *Eurographics Symposium on Point-Based Graphics*, 2007.

[18] **IDC**, Worldwide Big Data Technology and Services, *20122015 Forecast*, 1(2012), 233-485.

[19] **industriuscfo.com,** http://www.industriuscfo.com/7-benefits-using-big-data/, (accessed on February 2021).

[20] **Joao P.,** Topological data analysis and applications, *Institute Jozef Stefan*, 2017.

[21] **Julien T.,**Introduction to Topological Data Analysis *Sorbonne Universits, UPMC Univ Paris*, (2006), 1-22

[22] **lexico.com,** http://www.lexico.com/definition/ Powered by Oxford Dictionary (accessed on February 2021)

[23] **Luis B., Sergio M., Jos L.,** Deep Learning and Big Data in Healthcare: A Double Review for Critical Beginners, *Appl. Sci. 9*, 2019.

[24] **Lum P., Singh G., Lehman A., Ishkanov T., Vejdemo-Johansson M., Alagappan M., Carlsson J., Carlsson G.,** Extracting insights from the shape of complex data using topology, *Sci. Rep. 3*, 2013.

[25] **Oudot S.,** Persistence theory: From quiver representations to data analysis, *Am. Math. Soc.*, 1(2015), 209.

[26] **Perea J.,** A brief history of persistence, *Morfismos*, 23(2019), 1-16.

[27] **realpython.com** *https://realpython.com/what-can-i-do-with-python/*, (accessed on February 2021).

[28] **Robert J., Omer B., Matthew S., Eliran S., and Shmuel W.,** Persistent homology for random fields and complexes, *Borrowing Strength: Theory Powering Applications - A Festschrift for Lawrence D. Brown*, 6(2010), 124-143.

[29] **Robins V.,** Towards computing homology from finite approximations, *Topol. Proc.*, 24(1999), 503-532.

[30] **Saurabh J.,** Comprehensive Guide on t-SNE Algorithm with implementation in R and Python. *https://www.analyticsvidhya.com/blog/2017/01/t-sne-implementation-r-python/*, (2017) (accessed on February 2022).

[31] **Sidney A.,** Topology without Tears, *The MacTutor History of Mathematics Archive*, 217(2012).

[32] **Vclav S., Jana N., Fatos X., Leonard B.,** Geometrical and topological approaches to Big Data, *Future Generation Computer Systems*, 67(2017), 286-296.

[33] **yourdictionary.com,** https://www.yourdictionary.com/data-set-or-dataset (accessed on February 2021).