'THE END OF THEORY' AND THE TOPOLOGICAL DATA ANALYSIS IN THE BIOMEDICAL RESEARCH

ANDREI RODIN

Abstract.

1. Introduction

Topological Data Analysis (TDA) is a relatively recent method of Data Analysis based on the mathematical theory of Persistent Homology [36], [30], [14]. TDA proved effective in various fields of data-driven research including Life sciences and the Biomedical research. As the popular idiom goes, TDA helps to identify the *shape of data*, which turns out to be in many ways informative. But what precisely one can possibly learn from such shapes? How this method applies across different scientific disciplines and practical tasks? Sarita Rosenstock in her recent article [42]provided a very valuable presentation of TDA for general philosophical readership and explored the above epistemological questions in general terms. The present work extends Rosenstock's study in three different ways. First, it broadens the theoretical context of the discussion by bringing in some related epistemological problems concerning today's data analysis and data-driven research 2. Second, it brings the epistemological discussion on TDA into a wider historical context by pointing to some relevant earlier developments in the pure and applied mathematics 3, 4. Finally, the present Chapter focuses on applications of TDA in the Biomedical research and tests

Date: March 11, 2025.

theoretical epistemological conclusions obtained in the earlier sections of this work against some concrete examples 6.[1]

2. 'The End of Theory'

In his provocative 2008 popular article [3], Chris Anderson proclaimed the "end of theory". Anderson argued that given our dramatically increased capacities to collect, store, and proceed large volumes of digital data, such traditional methods of scientific research as modelling and theorising become "obsolete" and should be replaced by something akin to Google's algorithmic search. He argues that data sets of the "petabyte scale" don't admit for the usual treatment with traditional statistical methods but call "for an entirely different approach, one that requires us to lose the tether of data as something that can be visualized in its totality. It forces us to view data mathematically first and establish a context for it later." Anderson further argues that in the contemporary data-driven science the traditional concept of *causation* is no longer relevant, so that "correlation is enough" and should never be even tentatively interpreted in causal terms. On the positive side, Anderson's proposal amounts to letting "statistical algorithms find patterns where [the traditional] science cannot". As Rob Kitchin formulates Anderson's view "rather than testing a theory by analysing relevant data, new data analytics seek to gain insights *born from the data*" [27, p. 2].

Anderson justifies these and some other strong epistemological claims by referring to the success of Google's approach in the automated translation between natural languages, which unlike some other attempted approaches in the automated translation relies on statistical and syntactical but not semantical considerations. The recent success of Large Linguistic Models provides an additional support for Anderson's claims. A different motivation of Anderson's epistemological claims comes from the computational biology, namely, from the so-called "shotgun" DNA sequencing, which was invented in the late 1970s, and further advanced (using modern computers) and promoted by J. Craig Venter in the 2000s. Venter's approach is similarly based on a formal data analysis rather than testing some relevant theories and models.

In his harsh reply to Anderson Massimo Pigliucci claims that Anderson mistakes science for something different [38]. In Pigliucci's view, science is supposed to do exactly what Anderson's new science does not: to model, to theorise, to hypothesise and to explain. So Pigliucci argues that Anderson's notion of science simply does not deserve the name.

The reading of Anderson's arguments proposed in the present Chapter is equally critical but nevertheless more charitable. Pigliucci is right that modelling and theorising are valuable

[1] Since so many excellent presentations of TDA are found in the recent literature, which include informal presentations for the general academic readership [42], [30] as well as rigorous presentations for mathematicians and computer scientists [36], [14], the author of the present work focuses on epistemological features of TDA referring the reader to the aforementioned literature for learning about the mathematical details of the method itself.

epistemic procedures which should not be so easily given up. At the same time it would be wrong to dismiss Anderson's proposal only because it does'n square with the standard well-established ideas about science. Let us see what exactly Anderson proposes.

The identification of empirical patters stressed by Anderson as a distinctive feature of his proposed new scientific method equally plays a pivotal role in the traditional science. As remarks Frege in a different context, "[t]he discovery that the rising Sun is not new every morning, but always the same, was of very great consequence for astronomy." [15, p. 56]. One may add that the identification of the *pattern* of the "rising Sun" is perhaps even more important for science than solving Frege's problem of whether this pattern involves a single reappearing object or multiple objects. Thus the practice of identification of novel empirical patterns (including stable correlations) hardly diverges from the traditional forms of science (whether such an identification is achieved with a powerful computer and a smart algorithm or by a pair of non-assisted human eyes and a human brain).

What makes Anderson's proposal so controversial is his suggestion that in the new context of Big Data the identification of patterns not only remains necessary but also becomes *sufficient* in science, so more advanced epistemic tools like models and theories become "obsolete". There are two lines of argument in Anderson's paper that support this idea. First, he stresses that in the context of Big Data the identification of patterns by the naked eyes and the naked brain is impossible and, at the same time, it is possible and highly effective by using computers and appropriate algorithms. Second, he observes that in the context of Big Data such standard methodologies as theory testing and model-driven experimental design are unsuccessful. This leads Anderson to his pragmatic conclusion that the most effective way of acquiring a knowledge from the Big Data, namely, the algorithmic search for patterns, should be prioritised.

2.1. **Spurious correlations.** Ironically, the context of Big Data makes Anderson's idea that "correlation is enough" particularly prone to objection. Let us for the sake of argument accept after Anderson the Humean thesis according to which the concept of physical causation is unreliable and epistemically worthless. It turns out that the growing volumes of analysed data make correlations of variables computed on these data (using standard statistical methods) less informative anyway [8]. This effect, which can be formulated with a mathematical precision and explained via statistical considerations , gives rise to the so-called *spurious correlations* like that between the U.S. spendings on science, space and technology, and the number of people in U.S. suicided by hanging, strangulation and suffocation[2]. Such correlations are not informative not only because they are in odds with common ideas about causal relations but also because they are unspecific and can be fully explained away using only general statistical principles. Roughly, the spurious correlations are explained by the fact that the Big Data can be sufficiently big for making apparently improbable coincidences statistically probable.

[2] see https://www.datasciencecentral.com/spurious-correlations-15-examples/

The phenomenon of spurious correlations shows that at the "petabyte scale" (to use Anderson's word) traditional statistical methods developed in the 20th century become useless and need to be replaced by more appropriate methods of data analysis. This concerns among other things the standard methods of computing correlations. In his [3] erroneously assumes that the concept of correlation unlike that of causation is immune to the "deluge of data". It is not.

2.2. **Power of abstraction.** Anderson's example of gene sequencing by Venter which he offers as a paradigm of a new data-driven science is *not*, contrary to Anderson's claim, an example of scientific achievement made without modelling and theorising. Indeed, the concept of gene sequence is a highly developed theoretical concept making part of today's theory of biological reproduction and biological evolution. Venter's research could not be possibly done wholly outside of this theoretic framework. The "raw" data used by Venter in his research without such a framework could not be possibly collected. Thus Anderson's recommendation "view data mathematically first, and establish a context for it later" should be understood with a pinch of salt: a theoretical context matters already at the stage of data collection.

What the example of Venter's achievements demonstrates is not the irrelevance of modelling and theorising in today's data-driven science but rather the power of abstraction. When one wants to calculate the trajectory of a projectile using principles of Classical mechanics it is useful and even indispensable at certain point to forget about (i.e. abstract away) all the involved physical contents and proceed with purely mathematical (or even purely symbolic) manipulations; the obtained formal result is then (re)interpreted back in physical terms, and in this (interpreted) form constitutes the required solution. The effectiveness of this procedure demonstrates the power of mathematical abstraction across all sciences and the pure mathematics itself[3] The effectiveness of formal mathematical data analysis stressed by Anderson is of the same character. It demonstrates the power of formal methods and mathematical abstraction in various research fields where these methods are used carefully. But Anderson's examples do not support his radical thesis according to which raw data and clever mathematics is all what one needs in science.

2.3. **Pattern Identification.** Anderson's proposal to take the identification of patterns in empirical data seriously makes perfect sense even if one doesn't buy his claim that in the context of Big Data any further theorising is either impossible or not necessary [25]. The received historical narrative about science suggests thinking about pattern identification and pattern recognition is a primitive, archaic and mostly outdated forms of empirical research. The fact that the identification of observable patterns such as constellations is shared by

[3] Although pure mathematics can be described as *formal science* it admits for internal degrees of being formal. A simple example is given by Analytic Geometry where geometrical constructions are associated with certain algebraic equations. Given a geometrical problem one first formulates it in terms of appropriate equations, then solves the equations *formally*, i.e., forgetting about their geometric interpretation, and, finally, translates the obtained results back in the geometrical terms.

the science of Astronomy with the pseudo-science of Astrology suggests that the distinction between science and pseudo-science lies outside this primitive form of knowledge.

The fact that pattern identification is epistemically primitive does not mean, however, that it fully belongs to a remote historical past and cannot be both fruitful and problematic in today's science. Consider the case of Astronomy. Today's Astronomy is heavily laden by advanced astrophysical theories such as General Relativity (GR) and its heirs, and today's astronomical observations often aim at theory- and model-testing. The first empirical observation of gravitational waves [1], which were theoretically predicted by Albert Einstein one century earlier, exemplifies a success of the theory-focused approach in science. At the same time the 20th century Astronomy gives us some examples of important empirical discoveries that followed the more traditional path from the identification of certain unexpected and unexplained empirical patterns to accounting for and explaining away the discovered patterns in terms of appropriate theories. Such was the discovery of the *quasars*, which were identified in the early 1960s as unusual radio sources of unknown nature; later these sources were identified with active nuclei of remote galaxies [26]. It goes without saying that quasars cannot be observed by naked eye; the very notion of observation, which is relevant in the given case, is heavily theory-laden. Nevertheless the contrast with the case of gravitational waves is obvious: while the gravitational were first theoretically predicted and only later empirically discovered with an equipment built precisely for that purpose, the quasars were first empirically discovered and only later accounted for theoretically.

In research fields, which are not backed with advanced mathematically-laden theories like Astronomy and Astrophysics, the role of pattern identification is even more important. As we shall see in 6 this is the case of Biomedicine (among many other similar cases).

Thus, contrary to Anderson, we submit that there is no good reason to reject the notion of theory in data-driven research and give up all attempts to build theories using Big Data. But, this time in agreement with Anderson, we also submit that pattern identification is an important research activity which should not be ignored or underfunded. A stable reliable pattern has more epistemic value than an unreliable theory. But, other things being equal, a reliable theory still has more epistemic value than an identified empirical pattern. A further detail is that in the research practice the boundary between pattern identification and theory building is often blurred because the former is combined with proto-theoretical elements such as classifications and systematisations.[4]

3. Topological Data Analysis as a method of pattern identification

Let us now see how the pattern identification works in the TDA. The resulting patterns — otherwise referred to in TDA as *shapes* — are geometrically realised (usually low-dimensional) finite simplicial complexes [42]. Generally, TDA assigns to a given dataset

[4] See examples in 6.

more than just one shape: it continuously "zooms" the given data and picks up stable shapes at different scales wherever those are found.[5]

Since the TDA is whollyframedby the underlying theory of Persistent Homology it certainly does not support the idea of "data speaking of themselves free of human bias or framing" [27, p. 4]. But since this framing theory is purely mathematical it does not assume and does not tell anything about the nature of the analysed data. This is why TDA can be applied to any kind of data. Nevertheless the theory of Persistent Homology rigidly determines the class of possible shapes of those data that TDA can detect. Notice that unlike the traditional ways pattern identification and pattern recognition by naked eye (as in the aforementioned Frege's example), the TDA-based pattern identification and recognition is aided by a sophisticated mathematical theory, to wit, the theory of Persistent Homology. But since this mathematical theory is used in the TDA for purposes of general data analysis but not for modelling a specific natural or social phenomenon or representing some physical principle the resulting analysis remains formal; it allows one to detect a phenomenon but not, by itself, to obtaine its causal explanation.

It is instructive to compare an elaborated physical theory like GR with the mathematical theory of Persistent Homology as a part of the mathematical background of TDA. GR has a number of empirically testable consequences including the physical possibility of black holes and the existence of gravitational waves. An empirical test of such a hypothesis can be described in the epistemological terms as a very narrow (or focused) query, which needs highly specialised equipment (like that used in the LIGO experiment [1]), and which admits, at least ideally, only two possible responses: yes and no. By contrast, treating large volumes of empirical data with TDA can be described as a *distorted* (de-focused) query, which admits for a large (but nevertheless limited and pre-determined) spectrum of possible responses. When there is good theoretical reason to believe that collected data have some topological structure (as, for example, in the brain research where the brain structure is analysed of in terms of connections between neurones and groups of neurones [19] or in the case of information networks [2]) then the application of TDA can detect and specify this very structure. But TDA can be also applied *blindly* in the hope that detected topological shapes of analysed datasets would somehow carve nature in its joints and give one valuable insights into things and processes that have left their trace in the form of those datasets.[6]

The methodological question concerning these two kinds of queries — whether focused or rather distorted queries should be prioritised in science — is not new. Empirically-minded scientists and philosophers of the 17th and 18th centuries such as Francis Bacon (1561-1626) and Robert Boyle (1627 - 1691) rejected the medieval "Scholastic" tradition in the natural sciences which prioritised a logically refined theoretical speculation, in favour of a non-prejudged (but nevertheless systematic) data collection — very much in the spirit of

[5] "Zooming" should be understood here in terms of the global metrics defined on the given dataset, so the emerging topology is induced by this underlying metrics.

[6] Compare in 6 on "shallow" and "deep" applications of TDA.

the recent idea of "data speaking for themselves" . About the same time Galileo (1564-1642) pioneered the modern experimental method by combining experimental design with mathematical modelling of natural phenomena. Newton's *Principia* [24] are commonly seen as a crowning achievement of these developments, which provided a powerful paradigm for the modern science. Philosopher Immanuel Kant inspired by Newton's example, provided an elaborated account of the conceptual working of science, which justified the method of focused theoretically-laden experimental designed and devaluated a "blind" empirical search promoted by Bacon and his followers qualifying it as an archaic and outdated form of scientific study [17]. This didn't prevent, of course, scientists and philosophers from defending and pursuing alternative methodologies and research agendas during the following centuries.

What the lesson of TDA can teach us today about the above methodological dilemma? To begin with, it demonstrates that it is not a genuine dilemma. The choice is not between a blind and random empirical search, one the one hand, and testing elaborated theoretical hypotheses, on the other hand, but rather between more and less focused theoretical frameworks, which support larger and narrower empirical queries and, correspondingly, admit for larger and narrower spectra of possible empirical responses. Although the identification of a single empirical pattern is typically a more modest epistemic gain than the crucial testing of an important theory, using a weak multi-purpose theoretical framework such as TDA in an empirical study (rather than building a specific theory and designing experiments for its testing) is less costly. There is a tradeoff between the scope of a theoretical framework, the potential epistemic gain, and the research costs, which may also depend on some external factors. In any event there is indeed good reason to locate a part of available resources to "distracted" empirical studies using TDA and similar methods that give us a chance to discover new phenomena that lay wholly out of the scope of our current theories and then pursue their further study. In some case such empirical discoveries can be applied in practice before they are properly understood and explain (as this often happens in medicine).

4. Topological Data Analysis as a visualisation technique

In 1933 Hans Hahn published an influential article titled "The Crisis of Intuition"[7] where he argued contra Immanuel Kant that mathematical intuition[8] cannot play a foundational role in mathematics and science. In the support of his anti-Kantian view Hahn provided a number of mathematical and physical examples where, according to his analysis, Kant's account of how intuition and conceptual reasoning work together in mathematics and mathematically-laden science did not apply. In particular, Hahn pointed in this paper to geometrical constructions, which since the 1970s became known under the name of *fractals* because of their fractional Hausdorff dimension [32].

[7] [20], English translation [21]

[8] According to Kant, mathematical intuition splits into its spatial and temporal varieties, the former being related to geometry and the latter to arithmetic.

Referring to the graph of Weierstrass function (which is a fractal curve), Hans remarks that "the character of this curve entirely eludes intuition" [21, p. 84]. As a remedy of this alleged failure of mathematical intuition Hans points to a large program in the foundations of mathematics aiming at the "complete formalisation of mathematics", that is, a setting where "every new mathematical concept was to be introduced through a purely logical definition; every mathematical proof was to be carried through by strictly logical means" [21, p. 93].

A reason why fractals became so popular towards the end of the 20th century was, ironically, the booming development of computer imagery, which allowed Benoit Mandelbrot to visualise the geometrical construction, which Hahn assumed to be non-visualisable. Quite surprisingly these shapes looked even more "natural" than the more familiar Euclidean shapes, so Mandelbrot could use them for mimicking natural landscapes [32]. Whether or not Hahn was right urging in 1933 for "purely logical" foundations of mathematics he simply could not imagine how the Weierstrass curve would look like if its construction is continued up to the limits of human vision; perhaps he was not interested in studying this psychological question.

According to a popular historical narrative since the second half of the 19th century mathematics progressively developed toward higher levels of abstraction. Hahn's perceived *Crisis of Intuition* supports this narrative along with stories about the rise of the modern abstract algebra [47] during the first half of the 20th century, an extensive historical literature on the contemporary booming contemporary development of the axiomatic Set theory and the formal axiomatic viewpoint more generally [16], and the rise of Category theory during the second half of the same century [28], [41].[9]. Hahn's 1932 paper is just one historical evidence among so many that the development of mathematics towards higher abstraction and formalisation was indeed intended by its protagonists. It is impossible to deny that such efforts were successful. One of major protagonists of these developments, David Hilbert, expressed, however, a different view on the development of mathematics, which brings the receive narrative into question:

> In mathematics, as in any scientific research, we find two tendencies present. On the one hand, the tendency toward abstraction seeks to crystallize the logical relations inherent in the maze of material that is being studied, and to correlate the material in a systematic and orderly manner. On the other hand, the tendency toward intuitive understanding fosters a more immediate grasp of the objects one studies, a live rapport with them, so to speak, which stresses the concrete meaning of their relations.[10]

So in Hilbert's view a progress in abstract logical foundations of mathematics is wholly compatible with a progress in its intuitive representation and understanding. This is in

[9] In the 1950s Category theory was described as an "abstract nonsense" by one of its early developers and promoters Norman Steenrod, see [33]

[10] [22, Preface], quoted by [23]

spite of the fact that the two trends may compete against each other, and do so in the real history. The aforementioned case of fractals perfectly illustrates Hilbert's point: a concept that at certain point of history was conceived of as abstract and admitting no intuitive representation later acquired such a representation using a new type of imagery. This is a move from abstraction to *concretion*, not the other way round. Other similar examples are found in [40].

At the absence of any objective measure of the "level of abstractness", which can be applied to the whole of mathematics at any given moment of global historical time, the claim that "mathematics progressively becomes more abstract" can describe an epistemic priority or a vision but not a firm historical fact. Notice that the notion of the "whole" of mathematics is very vague and flexible. Should the elementary school mathematics around the world, the mathematics taught in engineering schools, the mathematics used in logistics and financial matters, etc. be included onto the global account? There is indeed a clear a sense in which the *Modern Algebra* as presented by van der Waerden [47] is more abstract than any kind of Algebra known before 1900. But how do we know that this development toward a higher level of abstraction was not compensated by the opposite move toward concretion elsewhere in mathematics?

4.1. **Cartesian imagery and Topological imagery.**

Whether the idea of an overall mathematical progress towards a higher abstraction can be saved or not, the above historical remarks show that the mathematical *concretion*, i.e., concrete forms of representation of mathematical concepts, deserves at least as much historical and epistemological consideration as mathematical abstraction. This is an appropriate starting point for an epistemological analysis of TDA, which brings along a new kind of mathematical imagery, which may complement and in some cases compete the more familiar Cartesian imagery. By the Cartesian imagery we understand here a representation of continuous (and typically smooth) functions using the rectangular Cartesian coordinates. The paradigm case here is the representation of trajectory of a point-like particle moving continuously through space (think of the parabolic trajectory of a projectile like a cannonball). Today's standard of mathematical literacy assumes a familiarity with this method of representation. It is taught in the secondary and undergraduate schools worldwide, and widely applied in economics, sociology, engineering, biology, physics, and many other fields[11]

[11] As the name indicates, the idea of Cartesian Coordinates dates back to the works of René Descartes in the 17th century. The Cartesian Coordinates achieved its modern form and became universally known, however, only much later. A history of Cartesian Coordinates, which focuses on their applications, diffusion and the imagery rather than on the conceptual development is still waiting to be written. It goes without saying that the conceptual part is very important: a graph of continuous function is not a mere image but an image interpreted in precise mathematical terms. Unless one is not familiar with the concept of continuous function that person cannot possibly interpret a given image as its graph. But that does not mean that the representation of functions in form of material images is somewhat trivial and doesn't deserve a historical study. As an example of a recent historical study that focuses on material and perceptional aspects of mathematics see [18]

TDA applies a different type of imagery, which is well-known to mathematicians (albeit in different visual forms) but so far remains largely unknown to the general public and even to many professionals who routinely use applied mathematics in their work; it enjoys today a social status similar to which the Cartesian imagery used to enjoy two or three centuries ago. Like the Cartesian imagery the TDA-related imagery is supported by a mathematical theory, which this time is a branch of Algebraic Topology called (the theory of) Persistent Homology. For that reason the TDA-related imagery can be called *topological.* If TDA fulfils its promises and replaces (or essentially complements) more traditional mathematical tools routinely used today for various practical purposes, the topological imagery and the associated topological conceptual optics will become a part of everyone's everyday life very much like graphs of continuous functions that represent changing inflation rates, average temperatures and other such things of common interest.

The history of Topology very well illustrates the conflicting co-existence of opposite trends of mathematical thought — one toward abstraction and the other toward concretion — noticed by Hilbert. The first systematic treatise on Topology by Johann B. Listing [31] which appeared in 1847, where the name of Topology first appeared publicly , is full of very concrete naturalistic examples. This includes a number of biological examples, which shows that the author followed his contemporary biological literature on morphology and made his own observation using microscope. In the very end of the 19th century Henri Poincaré published his fundamental work [39], which by today's standard is written in rather "concrete" because it uses a metrical set up of Riemanian geometry. Simultaneously begins the rise of highly abstract set-theoretic topology, pointed to by Hahn as a way out of the"crisis of intuition" [20]. In 1952 appears the fundamental monograph by Eilenberg and Steenrod [13] where the Algebraic Topology is presented axiomatically using Category theory. The language of category theory, on the one hand, can (and has been) been described as highly abstract but, on the other hand, it admits for concrete diagrammatic representation, which supports the topological intuition. The theory of Persistent Homology used in TDA developed during the early 1990s is an offspring of the above developments.

TDA applies topological concepts to an analysis of various empirical data. In this way TDA extends the scope of topological intuition far beyond the limits of pure mathematics. But what it can possibly tell us about the data and about things and processes, which have left their traces in the form of these data ?

5. Shallow and deep Topological Data Analysis in the Biomedical research

We shall try to answer the above question using applications of TDA in the Biomedical research as a case study. In a recent review article on that topic [44] the authors mention (among many other) the following achievements:

(1) the identification of a new type breast cancer tumour earlier undetected by other methods [35];

(2) the identification of two new subtypes of asthma [29];

(3) the visualisation of high-dimensional immunological data pertaining to co-infection of mice with influenza and pneumococcus, allowing for characterisation of distinct immune responses in various infection scenarios and stages within immune responses to co-infection [43];

(4) a highly accurate classification of subtypes of stomach lesions [12];

(5) a classification of type 2 diabetes mellitus (T2DM) disease trajectories [10];

(6) using molecular topological fingerprints (MTFs) for protein identification, classification, quantitative analysis of rigidity, and evolution of structural features during folding [48];

(7) the identification of distinct molecular states in the RNA hairpin folding process, providing insight into its mechanism [6];

(8) a study of the spread of contagions across a network from geometric and topological perspectives [45].

The following is a description and a tentative classification of cases (1)-(8) according to their epistemic character. (1)-(2) are cases of a pure pattern identification. The fact that the newly identified patterns of breast cancer tumour and of asthma were identified using TDA rather than a different method of data analysis is not supposed to have any effect on the obtained results: TDA proved effective in these case but the same new types of cancer tumour or of asthma could be in principle discovered by another method. TDA and its mathematical background (the theory of Persistent Homology) are not supposed to play any explanatory role in these cases. The role of TDA is limited here to the identification of new empirical patterns within a fixed theoretical framework. TDA is used in these cases as an external "optical instrument" (in the sense of conceptual optics) like a microscope or an X-ray machine: the principles of working of all those instruments have no specific relevance to things studied with these instruments. Accordingly, in these two cases the epistemic function of TDA reduces to pattern identification: TDA with its associated theory of Persistent Homology has no explanatory role in one's theoretical understanding of the breast cancer or asthma.

Case (3) is similar to (1)-(2) in that the theoretical background TDA is unrelated to the subject-matter (this time the subject matter is a set of scenarios of the immunological response to multiple infections) but the authors of [43] also stress the *representational* function of the TDA in their study. This example illustrates the above discussion on Cartesian and Topological imagery 4.1. The authors use both types of imagery: they use TDA with its supporting topological imagery in order to refine a Cartesian image that represents a theoretical prediction obtained earlier by the same authors using traditional mathematical modelling [43, Fig. 1]. This is how the authors describe their methodological view on using the TDA as an alternative to mathematical modelling:

> Motivated by the obvious potential of topological investigations in biomedical sciences, in the present study we seek to understand the evolution of the immune system as it responds to co-infection between virus and bacteria. Mathematical modelling research in influenza-pneumococcal co-infections has been a growing field within last years [...]. These previous approaches are based on differential equations constructed based on biological reasoning. While they are suitable tools to test different hypothesis, and have helped elucidate many of the details of the mechanisms of these intricate systems, these models are susceptible to bias by the designer and model complexity rapidly limits the reliability in the parameter fitting procedures [...]. In contrast, TDA is a tool that detects true patterns in the data, without imposing artificial assumptions. [43, p. 2]

Without trying to enter into the field of immunology where the author of the present work has no expertise, we notice that the above passage is objectionable from a purely methodological viewpoint. The idea of the TDA as a magic tool that allows one to "detect true patterns in the data" avoiding the usual human biases involved into the mathematical modelling and biological theorising (which makes an echo of Anderson's "end of theory" [3]) is, in our opinion, misleading. The TDA and the mathematical modelling with differential equation guided by a "biological reasoning" are not two different mathematical techniques of treating the available immunological data, which stand on equal footing and can be compared by their effectiveness. The two approaches have different epistemic goals. TDA helps to establish certain empirical *facts* via an analysis of relevant data and then to represent these facts in a concise and comprehensible manner. The mathematical modelling — and moreover a genuine scientific theorising combined with the mathematical modelling — has a different and higher epistemic ambition: it is supposed to elucidate and help us to *understand* the causal "mechanism" hidden behind the modelled phenomena. It goes without saying that scientific modelling and scientific theorising requires testing theories and models against empirical facts. Thus the TDA and the mathematical modelling do not really compete but rather complement each other: biological theories and related mathematical models can be checked against facts established with TDA (among other methods of data analysis).

Sasaki and her co-authors readily give up without noticing the traditional epistemic goals associated with scientific modelling and theorising and, without a further ado, limit their epistemic aim to establishing certain facts using TDA. It is undeniable that facts play a crucial role in all empirical sciences, and that modelling and theorising not tested against facts is a speculation but not a science. What is objectionable is the idea that establishing facts is the only goal that science can hope to achieve. When Sasaki and her co-authors refer to the "obvious potential of topological investigations in biomedical sciences" in the above quote, it appears that they talk about a potential *theoretical* significance of topological concepts in this area, i.e., to a possibility to build and develop biological *theories* using topological concepts. But TDA is not a theory but a method, its application in the biomedical research does not, by itself, introduce topological concepts into the biomedical theorising. Just like using an X-ray machine for studying a broken bone does not point to any functional role of X-rays in the human skeleton, the application of TDA for studying scenarios of the immune response does no, by itself, point to a potential role of topological concepts in our understanding of working of the immune system.

In scientific practice the distinction between theories and facts is often blurred. When a model has a lot of parameters that are supposed to be empirically adjusted, and when a fact is formulated using a heavy theoretical framework, the two may appear similar. It is hardly methodologically advantageous, however, to mix data, facts, theories and models completely. It is out of place to give here formal definitions to these notions but it is essential to distinguish them here informally. True, modelling and theorising requires a lot of human creativity. But it is misleading to describe this creativity in terms of "biases" as in the above quote. Facts, in their turn, typically depend on underlying theories. For example, the fact of the merger of black holes recorded in the LIGO experiment by detecting the gravitational waves issued as a result of this merger [1] could not be possibly established without General Relativity (GR) where the concepts of black hole and of gravitational waves are introduced. TDA is underpinned by the theory of Persistent Homology (PH) which equally required quite a lot of human inventiveness for its creation [37]. Unlike GR, PH is a mathematical but not physical theory. This implies that facts established with PH (via TDA) are *formal*; they reflect formal structures of data independently of where these data come from. TDA can tell us something important about scenarios of immune response (as in [43]) only when this method is used in a wider theoretical framework that includes some theory of the immune system and explains why the analysed data are relevant. The idea that TDA may present a viable alternative for modelling and theorising makes sense only in a situation where the concerned models and theories are sufficiently poor, so pursuing a less ambitious epistemic goal of establishing facts (laden with a sophisticated mathematical theory but not with any specific theory of the immune system) may be a good choice. But the idea that TDA provides a direct access to reality free of usual human biases is anyway wholly misleading.

As indicate their titles, in cases (4)-(5) TDA not only allows one to identify new patters but also suggests theoretical principles for their classification. In case (4) stomach lesions are classified according to their geometrical and topological properties, which are detected via the TDA of magnifying endoscopic images (ME) of the gastrointestinal tract. So in this case there is an understanding of how the topological "shapes of data" revealed by TDA translate into the shapes of stomach lesions identified with this method.

Case (6) is similar in this respect. Here the authors use the notion of "pseudo-time" that characterises the progression of a disease and "propose the joint use of pseudo time-series with TDA in order to illustrate the temporal characteristics of disease progression, so that disease trajectories can be constructed from the data using the topological model [of the data obtained with TDA] as a guide." [10, p. 2]. Once again, the path from the data to the wanted trajectories via TDA profiles aka the "topological model" or "shape" of the data in this case is transparent. Although the "topological model" referred to in the above quote is obviously not a full-fledged theoretical model of the disease in question, it is not just an external indication of certain state of affairs either[12]. The author's "topological model" does not directly represent but rather helpfully *encodes* the relevant features of analysed health records, which show how the studied disease may progress in the real time.

In case (7) the authors claim that TDA helps them to understand the "mechanism" of the RNA hairpin folding process. By the mechanism they understand here a detailed spatial description of the RNA hairpin folding, which involves topological concepts (such as a "contact" between different fragments of a RNA molecule). So TDA is, once again, used here for studying a natural process in which topological properties are essential.

In case (8) topological concepts are used for modelling the spread of contagions, and these very topological features are read off from the relevant empirical data using TDA. So in this case TDA is definitely used not for a mere pattern identification but for testing and an empirical adjustment of a theoretical mathematically-laden model of certain underlying natural process (viz. the spread of contagions) against the relevant data.

Thus cases (1)-(8) differ by their epistemological character. Cases (1)-(3) are examples of "shallow" applications of TDA. They are shallow in the sense that in these cases TDA has no special relevance to things and processes (namely, breast cancer, asthma and scenario of immune response) studied using this method. In cases (4)-(7) geometrical and topological concepts get involved (independently of TDA) into modelling of the phenomena studied with TDA. In these cases topological features of datasets revealed with TDA more directly reflect essential features of studied processes. Such applications of TDA can be described as being "deep". The deepest application of TDA in the above list is the case (8) where TDA is used for testing and the adjustment of a topological model of the spread of contagions.

Thus even if by default TDA is a method of formal data analysis, which may be applied to any set of data provided with a global metric, TDA can be also used in more specific contexts, which involves topological modelling and topological theorising. In this latter case using TDA is no longer presents itself as way to "allow data to speak for themselves" but is guided by appropriate theories, which share with this method a part of their mathematical background. This shows that Anderson's "end of theory" paradigm [3]is not only untenable (see 2.1) but also unnecessary because new effective methods of working with Big Data like TDA are compatible with the traditional modelling and theorising.

6. Conclusion

Since TDA is a specific method of data analysis, an epistemological study of TDA requires some preliminaries concerning the epistemological foundations of data science [11]. In the present work TDA has been framed into the continuing debate on the role of data analysis as a possible replacement for theory-building and theory-testing suggested back in 2008 by Anderson [3] 2, 3. Like many other critics of Anderson's radical proposal the author

[12] Compare cases (1)-(2).

of this work rejects it in its integrity but exploits its insights concerning the strengthened epistemic role of pattern identification and pattern identification in the context of Big Data combined with sophisticated mathematical methods of their analysis[13]. TDA clearly qualifies as such a method. Since TDA is underpinned by a mathematical theory of Persistent Homology (PH) this method cannot be described as theory-free; this theory provides a specific "conceptual optics" that allows for a specific interpretation of data using TDA. But the fact that PH is mathematical rather than physical or biological theory allows one to apply TDA to data of all types (provided that the data are appropriately prepared); this method is universal and not restricted to this or that domain of science.

The above case study of applications of TDA in the biomedical research adds important nuances to the above general picture. In some cases which have been described as *shallow* applications (cases (1)-(3) in 6) TDA is used as an external instrument that is wholly unspecific with respect to the studied biological processes. In such cases one can only hope that topological patterns identified using TDA in relevant datasets will reflect essential properties of the targeted biological entities rather than be mere artificially created figments[14]. But in other cases (cases (4)-(8) in 6) the application of TDA has a theoretically *deeper* character because there is some conceptual affinity between this particular method of data analysis and the background theories and models associated with the analysed data. Namely, in such cases topological concepts are applied not only as a part of TDA's mathematical background but also as a part of theoretical background related to the studied biological entities. In case (8) the studied process is the spread of contagions across a network where the network is described in geometrical and topological terms. In this case it is possible to read the relevant topological information from relevant data using TDA; this topological information adjusts the model to (and tests it against) empirical data just like this is done in the traditional mathematical modelling. There are also intermediate cases when TDA supports only some theoretical elements like a classification of the studied entities (cases (4)-(5)) but not a full-fledged theory or a model.

A remarkable feature of TDA, which partly explains its growing popularity, is its capacity to represent its outcomes in a visual form. Unlike the standard imagery used along with the traditional statistics and other methods of data analysis, TDA involves a novel type of imagery, which is tightly related to its mathematical background theory, to wit, the theory of Persistent Homology. In this Chapter the standard imagery has been called *Cartesian* (because of its historical roots), while the TDA-related imagery has been called *topological*. Even if using images for explaining the basic concepts of Algebraic Topology and Homological Algebra is by now a well-established educational practice producing such images automatically from datasets was not known before the rise of TDA. The relevance of TDA in many areas of empirical research where TDA proves useful allows scientists to explore and apply in their work many topological and homological intuitions, which until

[13] For an overview of various critical arguments against Anderson's theses of the "end of theory" see [11, section 6.2.].

[14] For a wider discussion on the epistemic role of mathematics in the context of data analysis see [34]

recently were known only to (and mastered only by) people having a special mathematical training.

Further, the present epistemological study of TDA demonstrates that the epistemological role of sensual (including visual) and intuitive representations of mathematical concepts is not limited to its educative and heuristic functions; they equally function as junctions that connect mathematical concepts to various forms of human experience that includes scientific experiments, observations and measurement.[15] During the last decades topological concepts have been successfully applied in various areas of science. This concerns theoretical physics (including Topological Quantum Field Theory [4]), chemistry [7], biology [46] and some other areas. As a general method of data analysis TDA does not exclusively belong to any of these fields; topological concepts are used in the TDA for analysing data of any type rather than for building some specific theories and models. Analysing data with TDA in the general case amounts to the identification of certain patterns by their topological properties that may reflect or not reflect essential features of studied things and processes. But a more attentive view on how TDA is used in practice shows that in the current research the distinction between pattern identification and theory- and model-building is often blurred .

It is instructive in this respect to compare the pioneering 1969 paper by René Thom who proposed to use topological ideas and topological mathematical apparatus for treating the biological phenomenon of *morphogenesis* [46] with the 2020 overview of topological approaches in biology by Ann S. Blevins and D.S. Bassett [5]. While Thom's approach can be described as a daring mathematically-laden scientific theorising aiming at a general biological phenomenon that doesn't have a satisfactory explanation, a large part of Blevins and Bassett's overview is reserved for a presentation of TDA that stresses its advantages in biology and explains the role of topology in data analysis. In the same article the authors also discuss the relevance of topological concepts in the accounting for signal flows in the complex biological systems, which is a clear case of mathematical modelling. The comparison suggests the following picture: while early applications of topology in biology like Thom's were theoretical and rather speculative — in the sense that they didn't have means for testing the proposed theories and models against empirical data — more recent applications were focused on the data analysis and some eventual modelling but avoided a systematic topological theorising. Anderson's "end of theory" [3] is a public expression of this latter trend. One may hope that in the future the existing gap between the theoretical speculation, on the one hand, and the study of empirical data, on the other hand, will

[15] Trying to save Kantian philosophy of mathematics in a new mathematical context that included Non-Euclidean geometries and other recent developments not taken into account by Kant himself, Ernst Cassirer in 1907 considered the effectiveness of mathematics in natural sciences to be the only relevant content of the concept of mathematical intuition: "The principle according to which our concepts should be sourced in intuitions means that they should be sourced in the mathematical physics and should prove effective in this field. Logical and mathematical concepts must no longer produce instruments for building a metaphysical "world of thought": their proper function and their proper application is only within the empirical science itself." [9, p. 44], the author's translation from German.

be filled with full-fledged topological theories and models in various domains of science, which will be testable against the available data using TDA or similar methods of data analysis.

References

[1] Abbott B.P. et al. (LIGO Scientific Collaboration, and Virgo Collaboration). Observation of gravitational waves from a binary black holes merger. *Physical Review Letters*, 116:061102, 2016.

[2] Aktas, M.E. et al. Persistence homology of networks: methods and applications. *Applied Network Science*, 4(61):1–28, 2019.

[3] Ch. Anderson. The end of theory: The data deluge makes the scientific method obsolete. *Wired*, June 2008.

[4] Birmingham D. et al. Topological field theory. *Physics Reports*, 209(4-5):129–340, 1991.

[5] A.S. Blevins and D.S. Bassett. *Topology in Biology*, pages 1–23. Springer International Publishing, Cham, 2020.

[6] Bowman, G.R. et al. Structural insight into rna hairpin folding intermediates. *Journal of American Chemical Society*, 130(30):96769678, 2008.

[7] Brown, D. Topological field theory. *Structural Chemistry*, 13(3-4):339–355, 2002.

[8] Calude, Ch.S. and Longo, G. The deluge of spurious correlations in big data. *Fundations of Science*, 22(3):595–612, 2017.

[9] Cassirer, E. Kant und die moderne mathematik. *Kant-Studien*, 12:1–40, 1907.

[10] Dagliati, A. et al. Using topological data analysis and pseudo time series to infer temporal phenotypes from electronic health records. *Artificial Intelligence in Medicine*, 108:101930, August 2020.

[11] Desai, J. et al. The epistemological foundations of data science: a critical review. *Synthese*, 200:469, November 2022.

[12] Dunaeva, O. et al. The classification of endoscopy images with persistent homology. *Pattern Recognition Letters*, 83(1):13–22, 2016.

[13] Eilenberg, S. and Steenrod, N. *Foundations of Algebraic Topology*. Princeton University Press, 1952.

[14] Baas, N.A. et al. (eds.). *Topological Data Analysis*. Springer, 2020.

[15] Frege, G. *On Sense and Reference in: Translations from the Philosophical Writings of Gottlob Frege, ed. by Geach and M. Black*, pages 56–78. Oxford: Basil Blackwell, 1952.

[16] Freudental, H. The main trends in the foundations of geometry in the 19th century. *Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress (Standford 1962)*, pages 613–621, 1960.

[17] Friedman, M. *Kant and the Exact Sciences*. Harvard University Press, 1992.

[18] Friedman, M. Paper, plaster, strings: Exploratory material mathematical models between the 1860s and 1930s. *Perspectives on Science*, 29(4):436–467, 2021.

[19] Gaputi, L. et al. Promises and pitfalls of topological data analysis for brain connectivity analysis. *NeuroImage*, 238:118245, 2021.

[20] Hahn, H. Die krise der anschauung. *Krise und Neuaufbau in den exakten Wissenschaften (Fünf Wiener Vorträge)*, 1933.

[21] Hahn, H. The crisis of intuition. *Hahn, H., Empiricism, Logic, and Mathematics: Philosophical Papers (Vienna Circle Collection v. 13)*, pages 73–102, 1980.

[22] Hilbert, D. and Cohn-Vossen, S. *Anschauliche Geometrie*. Springer: Berlin, 1932.

[23] Hilbert, D. and Cohn-Vossen, S. *Geometry and Imagination*. Chelsea Publishing(American Mathematical Society), 1952.

[24] Kant, Im. *Philosophiae Naturalis Principia Mathematica, translated into English by I.B. Cohen and A. Whitman*. California University Press, 1999.

[25] Kar, A.K. and Dwivedi, Y.K. Theory building with big data-driven research — Moving away from the "What" towards the "Why". *International Journal of Information Management*, 54:102205, 2020.

[26] Kellermann, K.I. The discovery of quasars. *Bulletin of the Astronomical Society of India*, 41:1–14, 2013.

[27] Kitchin, R. Big data, new epistemologies and paradigm shifts. *Big Data & Society*, pages 1–12, April-June 2014.

[28] Krömer, R. *Tool and Object: A History and Philosophy of Category Theory*. Basel: Birkhäuser, 2007.

[29] Kuo, C.H.S. et al. A transcriptome-driven analysis of epithelial brushings and bronchial biopsies to define asthma phenotypes. *u-BIOPRED*, 195(4):443–455, 2017.

[30] Lesnick, M. Studying the shape of data using topology. *The Institute Letter (IAS Princeton)*, Summer 2013.

[31] Listing, J.B. *Vorstudien der Topologie*. Cambridge University Press, 1847.

[32] Mandelbrot, B. *The Fractal Geometry of Nature*. W.H. Freeman and Company, 1982.

[33] McLarty, C. The uses and abuses of the history of topos theory. *The British Journal for the Philosophy of Science*, 41(3):351– 375, 1990.

[34] Napoletani, D. et al. The agnostic structure of data science methods. *Lato Sensu: Revue de la Socité de Philosophie des Sciences*, 8(22):44–57, 2021.

[35] Nicolau, M. et al. Topology based data analysis identies a subgroup of breast cancers with a unique mutational prole and excellent survival. *Proceedings of the National Academy of Sciences of the United States of America*, 108(17):7265–7270, 2011.

[36] Pascucci, V. et al. (eds.) *Topological Methods in Data Analysis and Visualisation*. Springer, 2011.

[37] Perea, J.A. A brief history of persistence. *Morfismos*, 23(1):1–16, 2019.

[38] Pigliucci, M. The end of theory in science? *European Molecular Biology Organisation Reports*, 10(6):534, 2009.

[39] Poincaré, H. Analysis situs. *Journal de l'École polytechnique*, 1:1–121, 1895.

[40] Rodin, A. How mathematical concepts get their bodies. *Topoi*, 29(1):53–60, 2010.

[41] Rodin, A. *Axiomatic Method and Category Theory (Synthese Library vol. 364)*. Springer, 2014.

[42] Rosenstock, S. Learning from the shape of data. *Philosophy of Science*, 88(5):1033–1044, 2021.

[43] Sasaki, K. et al. Topological data analysis to model the shape of immune responses during co-infections. *Communications in Nonlinear Science and Numerical Simulation*, 85:105228, 2020.

[44] Skaf, Y. and Laubenbacher, R. Topological data analysis in biomedicine: A review. *Journal of Biomedical Informatics*, page 104082, 2022.

[45] Taylor, D. et al. Topological data analysis of contagion maps for examining spreading processes on networks. *Nature Communications*, 6:7723, 2015.

[46] Thom, R. Topological models in biology. *Topology*, 8:313–335, 1969.

[47] Waerden, B. van der *Moderne Algebra*. Springer, 1930-1931.

[48] Xia, K and Wei, G.W. Persistent homology analysis of protein structure, flexibility, and folding. *International Journal for Numerical Methods in Biomedical Engineering*, 30(8):814–844, 2014.