# 1

# Topological Field Theory of Data: Mining Data Beyond Complex Networks

MARIO RASETTI AND EMANUELA MERELLI

## 1.1 A Philosophical Introduction

It has become increasingly obvious that very detailed, intricate interactions and interdependencies among and within large systems are often central to most of the important problems that science and society face. Distributed information technologies, neuroscience and genomics are just a few examples of rapidly emerging areas where very complex large-scale system interactions are viewed more and more as central to understanding, as well as to practical advances. Decision makers in these environments increasingly use computer models, simulation and data access resources to try to integrate and make sense of information and courses of action. There is also mounting concern that, in spite of the extended use of these simulations and models, we are repeatedly experiencing unexpected cascading systemic failures in society. We feel that, without resolving the issue of learning how to cope with complex situations, we also do not know enough about our methods of modeling complex systems to make effective decisions.

In the late eighties Saunders Mac Lane started a philosophical debate which, over thirty years later, is still going on with varying interest in the outcomes. This paper stems partly out of the crucial fundamental question that debate gave life to in contemporary science. This deep long-standing philosophical question, that can be formulated in several different ways, concerns mathematics. Are the formalisms of mathematics based on or derived from the facts and, if not, how are they derived? Alternatively, if mathematics is a purely formal game – an elaborate and tightly connected network of formal structures, axiom systems and connections – why do the formal conclusions in most of the cases fit the facts? Or, is mathematics invented or discovered? In the language of Karl Popper, statements of a science should be falsifiable by factual data; those of mathematics are not. Thus mathematics is not a science, it is something else. Yet the mathematical

network is tied to numberless sources in human activities, to crucial parts of human knowledge and, most especially, to the various sciences.

What is intriguing is not only the number of connections between mathematics and science, but the fact that they often bear on subjects which are at the very core of the mathematical network, not just on the basic topics at the edge of the network. The external connections of mathematics are numerous and tight, but they do not fully describe or determine the mathematical subjects. Basic mathematical concepts may be derived from human activity, but they are not themselves such activity; nor are they the phenomena involved as the background of such activity: the axiomatic method is a declaration of independence for mathematics.

Even though science has an inherent, natural tendency toward specialization, contemporary mathematics is more and more pursuing a general theory of structures. One such theory is category theory. "Category theory has come to occupy a central position in contemporary mathematics and theoretical computer science, and has also successfully entered physics. Roughly, it is a general mathematical theory of structures and of systems of structures. As category theory is still evolving, its functions are correspondingly developing, expanding and multiplying."[1] It is first of all a powerful language, a conceptual framework allowing us to see the universal components of a family of structures of a given kind, and how structures of different kinds are interrelated.

The message emerging is: the subjects of mathematics are extracted from the environment, that is from activities or phenomena of science and society. This notion of *extraction* is close to the more familiar term *abstraction*, with the intent that the mathematical subject resulting from an extraction is indeed abstract. Mathematics is not *about* human activity or phenomena, it is about the extraction and formalization of ideas and their manifold consequences. The formalization of such ideas in certain cases took centuries, but then it often opened the way to deep unexpected interconnections that in turn opened the way to looking at certain human activities in a completely new and diverse fashion.

The forces driving the development of the mathematical framework are manifold: for instance, generalization from specific cases, by analogy or by modification, and abstraction (once more) by analogy, or by deletion or yet by shift of focus; the appearance of novel problems; or simply, just plain curiosity. But questions arising from the variety of human and scientific activity have been and can be the most important sources of novel mathematics. Computer science also brings up new mathematical ideas. There is a wealth of new algorithms which bear on decisive conceptual aspects, such as the subtle question of computational complexity.

[1] Stanford Encyclopedia of Philosophy

On the other hand probably the most important fact in modern science is that dramatic change in paradigms that has seen reductionism challenged by holism. This is the story: an integrated set of methods and concepts have emerged in science since the mid-eighties under several designations, of which complexity science is the simplest and most comprehensive. Complex systems can be simply defined as systems composed of many non-identical elements, entangled in loops of nonlinear interactions. A typical example is neurons in the brain cortex. The challenge is to describe the collective properties of these systems, getting from the mere description of their components to the global properties of the whole system – in the example, from the description of neurons to the cognitive properties of the brain.

A difficult issue arises here, for when the composing elements and their interactions are highly simplified, the global properties are typically very hard to predict. The global description in terms of attractors of system model dynamics can be a strong and insightful simplification with respect to a full description of the *microscopic* components; this is exactly the same in thermodynamics, where global properties of a system can be described independently from the complete description of its microscopic elements, which is partially done, instead, by statistical mechanics. Yet, a real theory of complex systems, relating to the wide phenomenology of complex phenomena and data in the way in which statistical mechanics is related to thermodynamics, is still missing.

There is an overwhelming evidence that the current emphasis of numerous sciences, not only sciences of nature but sciences of society as well, on this novel paradigm of complexity (holism versus reductionism) sorely requires a rigorous scientific framing of its methodologies, which is not yet available. If it is true that wide classes of systems and problems from various disciplines share universal features that lead us to imagine the existence of common structures directing their dynamics, it is equally true that the simplified schemes whereby they are handled, once reduced to the conventional form of decision problems, can often be approached and solved only by resorting to very drastic, generally ad hoc simplifications. All problems dealt with in the framework of multi-agent complex systems, usually approached by network theory, belong to this latter family, which includes a huge number of applications, from bio- and eco-systems to economic and sociological decision making issues. Such simplifications are typically dictated by the utter lack of mathematical tools that are powerful or flexible enough to lead to a true theory.

A typical feature of complex systems is the *emergence* of nontrivial superstructures that cannot be reconstructed by a reductionist approach. Not only do higher emergent features of complex systems arise out of the lower level interactions, but the patterns that they create act back on those lower levels. This ensures that

complex systems possess a characteristic robustness with respect to large scale or multi-dimensional perturbations or disruptions, whereby they are endowed with an inherent ability to adapt or persist in a stable way. Because of their inherent structure, which requires analysis at many scales of space and time, complex systems face science with unprecedented challenges of observation, description and control. Complex systems do not have a *blueprint* and are perceived only through very large amounts of data. Therefore a typical task scientists are required to face is to simulate, model and control them, and mostly to develop theories for their behavior, control, management or prediction.

In science, methods generally come before theory; theory is the synthesis of knowledge gained by the application of systematic or heuristic methods. Although wide classes of systems from various disciplines share universal features that lead us to imagine the existence of common structures, their analysis is often based on drastic, generally ad hoc, simplifications, and their description resorts to the specific language proper to the most affine discipline, losing the richness of universality. On the other hand, a full theoretical understanding, for example, of the mechanism linking individual and collective behavior, along with the possibility of exploring the related systems with sufficiently powerful reliable simulations, cannot but lead to profound new insight in various areas. Metaphors should be avoided: metaphors are dangerous, because a metaphor is not a theory nor does it give much indication on specific applications.

To bridge the extraction of mathematical structures out of the phenomenology of complexity science and to give life to an efficient and complete collection of concepts and methods of mathematics appropriate for complexity theory is the challenge, and the most universal, potential setting frame for this is category theory: namely the construction of categorical structures for system modeling. Born with the aim of reorganizing algebra, looking not only at the objects (sets, groups, or rings) but also at the mapping between them (functions between sets, homomorphisms between groups or rings), category theory provides an elegant conceptual tool for expressing relationships across many branches of mathematics. It considers mathematical relations as *arrows* between *objects*. This approach fits in our case not only algebra but topology, where the arrows are continuous maps and objects are spaces, and geometry, with arrows that are smooth maps and objects which are manifolds. Category theory is a powerful, far-reaching formal tool for the investigation of concepts such as space, system, and even truth. It can be applied to the study of logical systems at the syntactic, proof-theoretic, and semantic levels. It is an alternative to set theory, with a foundational role for mathematics and computer science that answers many questions about mathematical ontology and epistemology.

Clearly, the choice to use the language of categories should not be made a priori, but should naturally impose itself due to the need to translate the seemingly purely mathematical objectives related to basic complexity science questions into theoretical computer science issues, and to establish a number of conceptual paradigms and technical instruments.

In complex systems, reconstruction is searching for a model that can be represented as a computer simulation program able to reproduce the observed data *reasonably well*. In this sense, reconstruction is the inverse problem of simulation. The statistics community addresses two closely related questions, namely, *what is a statistical model?* and *what is a parameter?* These questions, that are deeply ingrained in applied statistical work and reasonably well understood at an intuitive level as they are, are absent from most formal theories of modeling and inference. Whilst using category theory, these concepts can be well defined in algebraic terms, proving that a given model is a functor between appropriate categories. The objective that will guide us here is to construct an articulated and extended pathway connecting globally many apparently isolated (sub-)structures – those belonging to the functional (language) and behavioral (dynamics) features of complex systems; i.e., not simply gluing together a collection of local maps. This will be done by resorting to the language of category theory.

The novel approach to the problems of data-based complexity science described in this paper consists in the setting up of a new methodology, which is a sort of algebraic (in the sense of algebraic topology) complex systems theory, that pursues the idea that there exist suitable categories $\mathfrak{A}$ and $\mathfrak{B}$, functors $\mathcal{F} : \mathfrak{A} \to \mathfrak{B}$ and $\mathcal{G} : \mathfrak{B} \to \mathfrak{A}$, and a natural equivalence between them $\mathfrak{h} : \mathcal{F} \sim \mathcal{G}$, such that: $\mathcal{F}$ is a *simulation* and $\mathcal{G}$ is a *reconstruction*. In such schemes, systems of systems may be represented by $n$-categories, i.e., categories whose objects are arrows, arrows between arrows, and so on. Emergence may happen in any graph representing relationships between agents or multi-agents, in which spaces (or objects of some category) are attached to the vertices, and maps (or morphisms) are attached to the edges. As will be discussed in detail below, one can build out of such a graph an associated simplicial complex, whose *persistent homology* is the way to study its *shape* in a functorial way. Adaptivity arises in this way. Notice that no limitation is imposed in this perspective on the topology of the underlying graph, i.e., loops and self-loops are allowed, implying that systems with feedback can be included.

The categories to be involved in the conceptual scheme – why they emerge and how they can be linked together up to the completion of a global picture – come naturally out of the rationale of going beyond the traditional point of view and paradigms (networks, predicates, multi-agent schemes) by introducing in the framework of complex system theory the study of spaces in place of agents,

connecting them by morphisms instead of functions. Then one shall be able, from the study of the homology of the simplicial complexes generated by data clouds, to turn the data environment into a space of random variables connected by conditional probability distributions.

Categorification, the process of finding category-theoretic analogues of set-theoretic concepts by replacing sets with categories, functions with functors, and equations between functions with natural isomorphisms between functors satisfying the required *coherence laws*, can be iterated. This leads to $n$-categories, algebraic structures having objects, morphisms between objects, and also 2-morphisms between morphisms and so on up to $n$-morphisms. The morphisms of the old category *preserve* the additional structure.

This can be achieved through the description of the *process algebras* involved in terms of *quivers* and *path algebras*, and their representations. A quiver $\mathcal{Q}$ is a directed graph, possibly with self loops and multiple edges between two vertices. A representation of $\mathcal{Q}$ in a given category $\mathfrak{C}$ is obtained by attaching an object $\mathfrak{o} \in \mathfrak{C}$ to each vertex of $\mathcal{Q}$ and labeling each arrow of $\mathcal{Q}$ by a morphism between the objects sitting on its vertices. Given $\mathcal{Q}$ and $\mathfrak{C}$ there exists an algebra, $\mathcal{P}_{\mathcal{Q}}$, such that a representation of $\mathcal{Q}$ in $\mathfrak{C}$ is the same representation that would be obtained from $\mathcal{P}_{\mathcal{Q}}$ in $\mathfrak{C}$. Oriented paths in $\mathcal{Q}$ can be multiplied by concatenation and form a basis of $\mathcal{P}_{\mathcal{Q}}$. This gives an equivalence of categories and allows us to study the local properties of the quiver globally by means of its path algebra in a new scheme that is a very rich algebraic structure.

Graphical models, i.e., probabilistic models in which a graph describes the conditional independence structure between random variables, are commonly used in probability theory, statistics (particularly Bayesian statistics) and machine learning. The rules of discrete probability express the observed probabilities as polynomials in the parameters, parameterizing the graphical model as an algebraic variety. *Belief propagation*, Judea Pearl's algorithm, and all *message passing* methods of this kind are rooted in an environment of this sort. This work aims to overcome the limitations of these methods by importing the analysis tools from algebra, algebraic topology and quiver theory.

Homology is the mathematical device that converts information about a topological space into an algebraic structure in a functorial way. This implies that topologically equivalent (homotopic) spaces have algebraically equivalent (isomorphic) homology groups, and that topological maps between spaces induce algebraic maps (homomorphisms) on homology groups. Different homology theories have been developed for different spaces and needs; here we are interested in a special kind of homology which is called *persistent homology*. Given a discrete set in a higher dimensional space, persistent homology will allow us to attach to it a homological complex, which in turn will allow us to study the *shape* of the data set.

Long-lived topological features can thus be distinguished from short-lived ones in data sets, resorting to the simplicial complexes one can construct out of complex networks. The persistent homology of the complex identifies a graded module over a polynomial ring.

Most algebraic and combinatorial/configurational properties of the representation methods, such as structural isomorphism classes over graphs, maps and orders of local state evaluation, give rise to moduli over multi-graded vector spaces which are quiver representations. However, nearly all the usual homogeneity, symmetry and approximately infinite sizes that are essential for conventional statistical mechanics and other simplifications such as those necessary for the pursuit of network scaling and scale-free properties, are simply **not** present in meaningful treatments of interaction-based systems. The world of complex systems data is a much stranger, richer and more beautiful world than that. The challenge of understanding the collective emergent properties of these systems, from knowledge of components to global behavior is this: will Wigner's notion of "unreasonable effectiveness of mathematics" hold for complex systems as well?

Another deep philosophical question behind our work is an important one that was recently brought up by Vint Cerf [1]: whether or not there is any real *science* in computer science, namely if all the well posed questions can be approached by a truly scientific methodology: universal and self-contained. Of course, whenever computing implies the use of formal methods, i.e., mathematical techniques of some kind, it is reasonable to say that there is a rigorous element of science in the field. Computability, complexity analysis, theorem proving, correctness and completeness analysis, etc., are all abilities that fall into the category *scientific*. Since computing is a dynamical process rather than a static process, there is a need for stronger scientific tools that allow us to predict behaviors in computational processes. The challenge lies in being able to manage the explosive state space that arises from the interaction of the processes themselves with inputs, outputs, and with each other. In computer science, the need to constrain the unprecedented width of the state space range is often dealt with through the use of abstraction. Modeling is a form of abstraction, adequate to represent systems with fidelity, i.e., well defined in the abstract representation and suitable to be rigorously analyzed. Judea Pearl's *causal reasoning* in conditional probabilities is grounded on graphical models, linking the various conditional statements in chains of cause-effect: this introduces a sort of inherent time variable (reminding us of the *arrow of time* proper to statistical physics – the link being provided by entropy) and hence the ground for true dynamics.

Such a scenario is represented by diagrams analogous to those of Feynman's representation of quantum field interactions, that make it possible to construct

analytic equations that not only characterize the problem, but make its solution computable. Both are abstractions of complex processes, which aid our ability to analyze and make predictions about the system's behavior. Abstraction is a powerful tool: it eliminates unimportant details while revealing structure; a way of dealing with the problem that recalls statistical mechanics (smoothing out fluctuations, interaction-induced noise, renormalization) and chaos theory (the dynamical disorder effect of nonlinearity), where patterns emerge despite the apparent randomness of the processes. Our ability to understand and make predictions on data-represented complex processes rests on our cleverness in creating more efficient high-level query languages that allow unnecessary details to be suppressed and *theories* to emerge.

Information technology is facing its *fifth revolution*: the era of Big Data Science is challenged to handle information at unprecedented scales and needs to do so under diverse perspectives which share the common objective of selecting meaningful information from data. This means to be able to identify, within the space of data, the existing, typically hidden, correlation patterns, and formalize a consistent description of the data space structure that thus emerges. Such a structure contains the inherent, explicit representation of the organized information that data encode. Big Data Science needs to treat this massive corpus as a laboratory of the human condition. The challenge that arises is different, not only because it is much harder, but because – as the motto of complexity science asserts – *more is different*.

In this context, a 2008 editorial of *Wired* magazine with the provocative title "The End of Theory" prospected the idea that computers, algorithms and Big Data may generate more insightful, useful, accurate, true results than scientific theories, which traditionally rely on carefully crafted, targeted hypotheses and research strategies. This provocative notion has indeed entered not just the popular imagination, but also the research practices of corporations, governments and also academics. The idea is that data, shadow of information trails, can reveal secrets that we were once unable to extract, but that we now have the prowess to uncover, with no need of resorting to any underlying or pre-existing conceptual model.

Present work grows out of the conviction that, at today's scale, information is no longer a matter of simple low-dimensional taxonomy statistics and order, but rather of dimensionally agnostic pattern individuation. It calls for an entirely different approach; one that requires us to renounce the tether of data as something that can be embraced in its entirety. It instead forces us to view data mathematically, so as to be able to extract from it such rigorous information that will permit establishing its context. We claim that, contrary to the *Wired* magazine prophecy, this can be done and must be done, which establishes a well-defined theoretical context for a complex process that is unprecedentedly hard to handle. In other words, it is not

true that we no longer need to speculate and hypothesize, while simply we have to let machines lead us to patterns, trends, and relationships. We need to have a conceptual frame for handling the impending data deluge if we want to understand and control its implications, and construct a fully innovative theoretical conceptual structure that is a consistent stage for all plays.

On the other hand, a characteristic feature of complex systems is the *emergence* of nontrivial superstructures that cannot be reconstructed by a reductionist approach. Our goal is to build a tool for discovering directly from the observation of data those mathematical relations (patterns) that emerge as correlations among events at a global level, or alternatively, as local interactions among systemic components. Not only do higher emergent features of complex systems arise out of such lower level interactions, the patterns they create may also react back, implying the capacity to develop tools to support a learning process as well.

We develop here a topological field theory for data space, a concrete (though conceptual) objective that is itself proof-of-concept of its breakthrough capacity. The problem at stake can be seen as a far-reaching evolution/generalization of *data mining*, which is the analysis step of knowledge generation in data sets, and focuses on the discovery of unknown features that data can conceal. Data mining uses typically artificial intelligence methods (such as *machine learning*), but often with different goals. Machine learning employs *unsupervised learning* to improve the learner accuracy in the design of algorithms, allowing computers to evolve its major focus: to recognize complex patterns in data and make intelligent decisions based on it. The difficulty here is that the set of all possible behaviors, given all possible inputs, is too large to be covered by the set of observed examples (training data). Predictions are based on known properties learned from the training data: the true task of data mining is then the automatic analysis of large quantities of data, aimed at extracting interesting patterns to be used in predictive analytics. We argue that the data tsunami we are facing can be dealt with only by mathematical tools that are able to incorporate data in a topological setting, enabling us to explore the space of data **globally**, so as to be able to control its structure and hidden information.

In spite of their robustness – namely the capacity they are endowed with to adapt and persist in stable forms – and the emphasis of science on the paradigm of complexity, complex systems are hard to represent and harder to predict. One of the reasons for this is that complex systems knowledge is mostly based not on a shared, well-defined phenomenology, but on data. Yet there are clear elements of universality in the dynamical features of such systems. A real theory of complex systems having a direct bearing on complex phenomena and data in the same way as statistical mechanics bears on thermodynamics, is still not available. A deeper question is thus: can it ever be available? Gödel's theorem and Cantor's set theory

appear to forbid it, implying as they do that an infinite multiplicity of conceptual models should exist, but the challenge of a *statistical dynamics* with no background ergodic hypothesis, no thermodynamic limit, no identical *particles* (agents), and above all, not based on repeatable experiments but data driven, is certainly there and needs to be faced. The latter reason is what makes us focus our attention first on the Big Data issue.

Data collection, maintenance and access are central to all crucial issues of society, because the increasingly large influx of data bears not only on science but on a correct governance of all societal processes as well. Large integrated data sets can potentially provide a much deeper understanding of nature but they are also critical for addressing key problems of society. We claim that the data tsunami we are facing can be dealt with only with mathematical tools that are able to incorporate data information in a geometric/topological way, based on a space of data thought of as a collection of finite samples taken from (possibly noisy) geometric objects.

Our work rests on three pillars, interlaced in such a way as to reach the specific objective of devising a new method to recognize structural patterns in large data sets, which allows us to perform data mining in a more efficient way and to extract more easily valuable effectual information. Such pillars are: i) topological data analysis (homology driven), and the related geometric/algebraic/combinatorial architecture; ii) topological field theory for data space as generated by the (simplicial complex) data structure, the construction of a measure over data space, and the identification of a gauge group; iii) formal language (semantic) representation of the transformations presiding the field evolution.

## 1.2 The Reference Landscape

Complex Systems are ubiquitous: they are complex, multi-level, multi-scale systems and are found everywhere in nature and also in the Internet, the brain, the climate, the spread of pandemics, in economy and finance; in other words, in society. Here we intend to address the deep, intriguing question that has been raised in a previous section about complex systems: can we envisage the construction of a bona fide *Complexity Science Theory*? In other words, does it make sense to think of a conceptual construct playing for complex systems the same role that Statistical Mechanics played for Thermodynamics?

As it has already been mentioned, the challenge is indeed enormous. In statistical mechanics a number of assumptions play a crucial constraining role: i) *ergodicity*, ensuring that all accessible states of the system considered are visited with equal probability; ii) the so called *thermodynamic limit*, $N \to \infty$, requiring

that the number of degrees of freedom $N$ (proportional to the number of particles, measured essentially by the Avogadro number), could be assumed as essentially infinite; iii) particles are identical (or possibly indistinguishable): particles of the same species are identical and interact with each other pairwise all in the same way, that is, obeying the same interaction law – in the quantum case they are indistinguishable; iv) an analytical structure is definable for the underlying dynamics, namely equations of motion exist at the micro-scale – analyticity breaking and singularities only appear as a signal of the macro-phenomenon of phase transition; v) experiment-based – phenomenology, implying that phenomena are repeatable, as in reductionist science: under the same initial and boundary conditions the same experiment must give the same outcome.

In contrast, typically complex systems, in particular those representing societal phenomena, have the following hallmarks: i) they are NOT ergodic; ii) their number of agents, $N$, is ordinarily finite, even though it can be large on a social scale; iii) their agents are NOT identical – they are quite distinguishable complex systems themselves, with their strategies and autonomous behaviors; iv) they are NEVER representable by analytic, perhaps in certain cases not even by recursive, functions; v) above all, they are DATA-based, usually NO repeatable experiment is possible under external control.

The world we live in is no doubt complex and dramatically data-based. More than 4 billion people (more than half of the world's population) own a mobile phone (which makes this the first device in human history owned by more than a half of the world's inhabitants); every day over 300 billion e-mails and 25 billion SMSs are exchanged, 500 million pictures are uploaded on Facebook, etc. The information created and exchanged in a year added up in 2013 to 4 zettabytes (1 zettabyte $= 10^{21}$ bytes) and every year it grows 40% (in 4 years it will reach a yottabyte, $10^{24}$, a number larger than Avogadro's number!). For this reason we concentrate first on the last item of the above list: *data*, indeed Big Data. The challenge is to extract all the information, as a norm hidden within, from the huge collection of data flowing in and around complex systems.

Big Data have a variety of diverse features. They have always been present in science where they have played a central role (though today even science has difficulties in dealing with the immense quantity of data made available by measures and experiments; see, e.g., the Hubble and Genome projects, the CERN data archive, etc.): typically scientific data is well organized in high quality data-bases. Today Big Data also plays a role in society, where it may for the first time allow for a true societal tomography, making possible predictions not envisioned before (see, e.g., the H1N1 pandemics of 2009), or for unprecedented targets and strategies. Big Data pose a demanding hardware challenge, (high performance computing), and also a strenuous data manipulation challenge, both

in computer science (new computing paradigms; interaction-based computing; beyond the Turing machine) and data analytics (new approach to data mining; nonlinear causal inference). Also, the ever-more blurred boundaries between the digital and physical worlds that characterize our digitalized global world are bound to progressively fade away as IT becomes an integral part of the fabric of nature and society.

A parallel goal is to endow IT with innovation and to use more and more efficient tools to play its role in the hard process of turning *data* into *information*, information into *knowledge*, and eventually knowledge into *wisdom*; in other words, to give life to a new paradigm for data manipulation capable of managing the complex dialectic relations between structural and functional properties of systems, in a way analogous to that with which the human brain interacts with information and behaves as a set of embodied computers. An exercise in *artificial intelligence*.

We explore the possibility of taming Big Data with topology (the geometry of *shapes*), building on a fundamental notion from computer science when dealing with data: the concept of *space of data*. It is the latter that provides the structure (represented geometrically) within which information is encoded, such as the frameworks for algorithmic (digital) thinking, and the lode in which to perform data mining, i.e., to extract patterns of correlated information. It is the very notion of data space that engenders the objective: finding new ways – based on its geometrical (topological) and combinatorial features – to extract (*mine*) information from data.

The ideas proposed by Carlsson, Edelsbrunner and others will now be expanded upon. They all argued that geometry and topology are the natural tools to handle large, high-dimensional, complex spaces of data in this process. *Why*? Because *global*, though *qualitative*, information is relevant; data users aim to obtain maximum knowledge, i.e., to understand how data is organized on a large, global scale rather than locally. *Metrics are not theoretically justified*: while in physics, most phenomena naturally lead to elegant, clear-cut theories which imply – as an outcome of the theory itself – the metrics to be used; in the life or social sciences this is either less cogent or it is simply not there. *Coordinates are not natural*: data is typically conveyed and received in the form of strings of symbols, typically numbers in some field, and vector-like objects whose *components* have no meaning as such and whose linear combinations are not objects in data space. In other words, the space of data is *not* a vector space. Thus those properties of data space that depend on a specific choice of coordinates cannot be considered relevant. *Summaries only are valuable*. The conventional method of handling data is based on the construction of a graph (*network*) whose vertex set is the collection of points in data space (each point possibly itself a collection of data) where two vertices are connected by an edge if their *proximity measure* is, say, less than a

given threshold $\eta$; followed by the attempt to find (determine) the optimal choice of $\eta$. The complete diagram that illustrates the arrangement produced by data hierarchical clustering is, however, much more informative. It is able to capture at once the summary of all relevant features with all possible values of $\eta$. The difficulty is to get to know how the global features of data space vary upon varying $\eta$.

For all these reasons the methods to be adopted should be inspired by topology, because: *topology* is the branch of mathematics that deals with both *local* and *global qualitative* geometric information in a (topological) ambient space, specifically connectivity, classification of loops and higher dimensional manifolds, and invariants, which are properties that are preserved under homeomorphisms of the background space. *Topology* studies geometric properties in a way that is *insensitive to metrics*; it ignores the notion of distance and replaces it just with the concept of proximity ($\eta \approx$ connective *nearness*, in the sense of Grothendieck topology). *Topology* deals with those properties of geometric objects that *do not depend on coordinates* but only on intrinsic geometric features; it is *coordinate-free*.

Besides, and perhaps even more importantly, in topology relationships involve maps between objects; thus they are naturally a manifestation of *functoriality*. Also, the invariants are related not just to objects, but to maps between objects as well. Thus functoriality reflects an inherent *categorical* structure, allowing for computation of global invariants from local information.

Finally, the whole information about topological spaces is typically faithfully contained in their *simplicial complex* representation, that is itself a piece-wise linear (PL), combinatorially complete, discrete realization of functoriality. As already mentioned, the conventional way to convert a collection of points in data space into a global object is to use the vertex set of a network, whose edges are determined by proximity. However, while such a graph is able to capture data connectivity (a local property of the network), it ignores a wealth of higher order global features, which are instead well discerned by a higher-dimensional object, the *simplicial complex*, that can be thought of as the scaffold (the 1-skeleton) of the graph. The latter is a PL space built by gluing together simple pieces (the simplices) identified combinatorially along their faces, which are obtained by the completion of the graph.

## 1.3 The Challenge

Current thinking in Information Technology is at the crossroads of different evolution pathways. On the one hand, is what is now universally referred to as the

*Big Data* question [2], which urges fully innovative methodologies to approach data analytics – in particular data mining – to be able to extract information from data with the required efficiency and reliability. On the other hand, is the ever increasing number of real world instances, in science as well as in society, of problems that ask us to go computationally *beyond Turing* [3], which touches on such basic issues as *decidability*, *computability*, and even *embodied computation*.

Following Cerf's view, we assume modeling of computational processes as the most inspiring candidate for the construction of a true *theory* that is able to lead to credible predictions about complex processes through the analysis of the large data sets that represent them. We also keep Pearl's diagrams [4], whose surprising analogy with Feynman's representation of interactions in quantum field theory, with cause-effect relations replacing time flow direction, which was already previously mentioned, in mind as a reference paradigm for the construction of analytic equations that not only fully characterize this type of problem, but make their solution accessible. A crucial issue here is indeed to eliminate unimportant details while revealing the relevant underlying structure: a method well known to statistical mechanics (this is exactly what the renormalization group method does), dealing with fluctuations and noise induced by interactions, and to chaos theory (one of the dramatic dynamical effects of nonlinearity), where patterns emerge despite the apparent randomness of the process.

This paper intends to describe a long-term program designed to generate a novel pathway to face some of the challenges posed above – in particular, the issue of sustaining predictions about the dynamics of complex processes through the analysis of Big Data sets – by paving the way for the creation of new high-level query languages that allow insignificant details to be suppressed and meaningful information to emerge as *mined out* correlations. The main goal of this paper is to find the definition of a theoretical framework, described essentially as a nonlinear topological field theory, as a possible alternative to conventional machine learning or other artificial intelligence data mining techniques, allowing for an efficient analysis of and extraction of information from large sets of data.

The approach proposed differs from all previous ones in its deep roots in the inference of globally rather than locally coded data features. Its focus is on the integration of the pre-eminent constructive elements of topological data analysis (*facts* as *forms*) into a topological field theory for the data space (which becomes in this way the logical space of forms), relying on the structural and syntactical features generated by the formal language whereby the transformation properties of the space of data are faithfully represented. The latter is a sort of *language of forms* recognized by an automaton naturally associated with the field theory.

This perspective has a profound, far-reaching philosophical meaning. As Wittgenstein beautifully phrased it [5]: "The world is the totality of facts, not

things. … The facts in logical space are the world. … A logical picture of the facts is a thought." and "To imagine a language means to imagine a form of life. … The meaning of a word is its use in the language game."

The decisive outcome of the approach proposed will be a way to extract directly from the space of observations (the collection of data) those relations that encode – by means of this novel language – the emergent features of the complex systems represented by data; *patterns* that data themselves describe as correlations among events at the global level, the result of interactions among systemic components at local level. The complex system's global properties are hard to represent and even harder to predict, just because – contrary to what happens in traditional reductionist science – complex systems knowledge in general does not bear on repeatable experiments and phenomenology that, incidentally, provide the necessary shared information leading to the statistical characteristics of the system properties, but on *data* or on virtual artificial representations of real systems built out of data.

There are three bodies of knowledge that constitute the three pillars our scheme rests on, which need to operate synergically: i) *homology theory*; ii) *topological field theory* and iii) *formal language theory*. The *singular homology methods* (i) furnishes the necessary tools for the efficient (re-)construction of the (simplicial) topological structures in the space of data which encode patterns. It enables us to make topological data analysis homology driven and coherently consistent with the global topological, algebraic and combinatorial architectural features of the space of data, when equipped with an appropriate *measure*. *The topological field theory* (ii) provides the construct, mimicking physical field theories (as connected to statistical field theories), for extracting the necessary information to characterize the patterns in a way that might generate, in view of the field nonlinearity and self-interaction, the reorganization of the data set itself, as feedback. The construction of the *statistical/topological field theory of data space*, is generated by the simplicial structure underlying data space, by an action and the corresponding fiber (block) bundle. An action depends on the topology of the space of data and on the nature of the data, as they characterized by the properties of the processes whereby they can be manipulated, a *gauge group* that embodies these same two features: data space topology and process algebra structure. The *formal language theory* (FLT) (iii) offers the way to study the syntactical aspects of languages generated by the field theory through its algebraic structure, i.e., the inner configuration of its patterns, and to reason and understand how they behave. It allows us to map the semantics of the transformations implied by the nonlinear field dynamics into automated self-organized learning processes. These three pillars are interlaced in such a way as to allow us to identify structural patterns in large data sets, and efficiently perform data mining. The outcome is a new

pattern discovery method, based on extracting information from field correlations that produces an automaton as a recognizer of the data language.

### 1.4 Step One: Topological Data Analysis

The main pillar of the construction of our theory is the notion of data space, whose crucial feature is that it is neither a metric space nor a vector space – a property that is unfortunately still uncritically assumed even by the most distinguished authors (see, e.g., Hopcroft and Kannan [6]) – but it is a topological space. This is at the root of most aspects of the scheme proposed: whether the higher dimensional, global structures encoding relevant information can be efficiently inferred from lower dimensional, local representations; whether the reduction process performed (filtration; the progressive finer and finer simplicial complex representation of the data space) may be implemented in such a way as to preserve maximal information about the global structure of data space; whether the process can be carried over in a truly metric-free way [7]; whether from such global topological information *knowledge* can be extracted, as well as correlated information, in the form of patterns in the data set.

The basic principles of this approach stem from the seminal work of a number of authors: G. Carlsson [8], H. Edelsbrunner and J. Harer [9], A. J. Zomorodian [10], and others. Its fundamental goal is to overcome the conventional method of converting the collection of points in data space into a *network* – a graph $\mathcal{G}$ encompassing all relevant *local* topological features, whose edges are determined by the given notion of *proximity*, characterized by parameter $\eta$ that fixes a coordinate-free metric for *distance*. Indeed, while $\mathcal{G}$ captures pretty well *local* connectivity data, it ignores an abundance of higher-order features, most of which have a *global* nature, and misses its rich and complex combinatorial structure. All these can instead be accurately perceived and captured by focusing on a different object than $\mathcal{G}$, say $\mathcal{S}$. $\mathcal{S}$ is a higher-dimensional, discrete object, of which $\mathcal{G}$ is the 1-skeleton, generated by combinatorially completing the graph $\mathcal{G}$ to a *simplicial complex*. $\mathcal{S}$ is constructed from higher and higher-dimensional simple pieces (simplices) identified combinatorially along their faces. It is this recursive and combinatorially exhaustive way of construction that makes the subtlest features of the data set, seen as a topological space $X \sim \mathcal{S}$, manifest and accessible.

In this representation, $X$ has an *hypergraph* structure whose hyperedges generate, for a given $\eta$, the set of relations induced by $\eta$ itself as a measure of proximity. In other words, each hyperedge is a *many-body relational* simplex, namely a simplicial complex built by gluing together lower-dimensional relational simplices that satisfy the $\eta$ property. This makes $\eta$ effectively metric independent:

in fact an $n$-relation here is nothing but a subset of $n$ related data points, satisfying the property represented by $\eta$. Dealing with the simplicial complex representation of $X$ by the methods of algebraic topology, specifically the theory of persistent homology that explores it at various proximity levels by varying $\eta$, i.e., filtering relations by their robustness with respect to $\eta$, allows for the construction of a parameterized ensemble of inequivalent representations of $X$. The filtration process identifies those topological features which persist over a significant parameter range, making them eligible as candidates to be thought about as *signal*, whereas those that are short-lived can be assumed to characterize *noise*. Moreover, it implicitly defines the notion of an $\eta$-parametrized semigroup connecting spaces in the ensemble.

Key ingredients of this form of analysis are the homology groups, $H_i(X)$, $i = 0, 1, \ldots$, of $X$ and in particular the associated *Betti numbers* $b_i$, the $i$-th Betti number, $b_i = b_i(X)$, being the rank of $H_i(X)$ – a basic set of topological invariants of $X$. Intuitively, homology groups are functional algebraic tools that are easy to deal with (as they are abelian) to pick up the qualitative features of a topological space represented by a simplicial complex. They are connected with the existence of *i-holes* (holes in $i$ dimensions) in $X$. Holes simply mean $i$-dimensional cycles which don't arise as boundaries of $(i+1)$ or higher-dimensional objects. Indeed, the number of $i$-dimensional holes is $b_i$, the dimension of $H_i(X)$, because $H_i(X)$ is realized as the quotient vector space of the group of $i$-cycles with the group of $i$-boundaries. In the torsion-free case, knowing the $b_i$'s is equivalent to knowing the full space homology and the $b_i$ are suffcient to fully identify $X$ as topological space.

Efficient algorithms are known for the computation of homology groups [11]. Indeed, for $\mathcal{S}$, a simplicial complex of vertex-set $\{v_0, \ldots, v_N\}$, a simplicial $k$-chain is a finite formal sum $\sum_{i=1}^{N} c_i \sigma_i$, where each $c_i$ is an integer and $\sigma_i$ is an oriented $k$-simplex $\in \mathcal{S}$. One can define on $\mathcal{S}$ the group of $k$-chains $\mathcal{C}_k$ as the free abelian group which has a basis in one-to-one correspondence with the set of $k$-simplices in $\mathcal{S}$. The boundary operator

$$\partial_k : \mathcal{C}_k \to \mathcal{C}_{k-1} \tag{1.1}$$

is the homomorphism defined by:

$$\partial_k \sigma = \sum_{i=0}^{k} (-1)^i (v_0, \ldots, \widehat{v_i}, \ldots, v_k), \tag{1.2}$$

where the oriented simplex $(v_0, \ldots, \widehat{v_i}, \ldots, v_k)$ is the $i$-th face of $\sigma$ obtained by deleting its $i$-th vertex.

In $\mathcal{C}_k$ elements of the subgroup

$$Z_k = \ker(\partial_k) \tag{1.3}$$

are referred to as *cycles*, whereas those of the subgroup

$$B_k = \mathrm{im}(\partial_{k+1}) \tag{1.4}$$

are called *boundaries*.

Direct computation shows that $\partial^2 = 0$, simply meaning that the boundary of anything has no boundary. The abelian groups $(\mathcal{C}_k, \partial_k)$ form a *chain complex* in which both $B_k$ and $Z_k$ are contained; $B_k$ is included in $Z_k$.

The $k$-th homology group $H_k$ of $\mathcal{S}$ is defined to be the quotient abelian group

$$H_k(\mathcal{S}) = Z_k / B_k. \tag{1.5}$$

There follows that the homology group $H_k(\mathcal{S})$ is non-zero exactly when there are $k$-cycles on $\mathcal{S}$ which are not boundaries, meaning that there are $k$-dimensional holes in the complex.

Holes can be of different dimensions. The rank of the $k$-th homology group, the number

$$b_k = \mathrm{rank}(H_k(\mathcal{S})), \tag{1.6}$$

the $k$-th Betti number of $\mathcal{S}$, gives just a measure of the number of $k$-dimensional holes in $\mathcal{S}$.

Persistent homology is generated recursively, starting with a specific complex $\mathcal{S}_0$, characterized by a given $\eta = \eta_0$ and constructing from it the succession of chain complexes $\mathcal{S}_\eta$ and chain maps for an increasing sequence of values of $\eta$, say $\eta_0 \leq \eta \leq \eta_0 + \Lambda$, for some $\Lambda$. The size of the $\mathcal{S}_\eta$ grows monotonically with $\eta$, thus the chain maps generated by the filtration process can be naturally identified with a sequence of successive inclusions.

In algebraic topology most invariants are difficult to compute efficiently, but homology is not: it is actually somewhat exceptional not only because – as we have seen – its invariants arise as quotients of finite-dimensional spaces but also because some of its properties can sometimes be derived from *physical* models. In standard topology, invariants were historically *constructed* out of geometric properties and manifestly able to distinguish between objects of different shape but homeomorphically invariant globally. Other invariants were instead obtained in physics, and these were in fact *discovered*, based, e.g., on topological quantum field theory technology [12]. These invariants provide information about properties that are purely topological but one cannot detect, not even guess, based only on geometric representation.

It is this perspective that we adopt here, namely the idea of constructing a reliable *physical* scenario for data spaces, where no structure is visible. *Physical* should of course be interpreted metaphorically: we aim to construct a coherent formal framework in the abstract space of data, where no equation is available giving the information it encodes as an outcome, that is capable to describe through its topology the hidden correlation patterns that link data into information. This is metaphorically analogous to what one has, say, in general relativity, when a given distribution of masses returns the full geometry of space-time. Here we expect that a given amount of information hidden in data would return the full topology of data space. Of course we don't have a priori equations to rely on, yet we argue that a topological, nonlinear field theory can be designed over data space whereby global, topology-related pattern structures can indeed be reconstructed, providing a key to the information they encode.

All this bears of course on how patterns must be interpreted, as it deals rather with pattern *discovery* than pattern *recognition*. This requires at least a remark. In logic there are approaches to the notion of pattern that, drawing on abstract algebra and on the theory of relations in formal languages – as opposed to others that deal with patterns via the theory of algorithms and effective constructive procedures – define a pattern as that kind of structural regularity, namely organization of configurations or regularity, that one identifies with the notion of *correlations* in (statistical) physics [13]. These logical paradigms will guide our strategy.

A subtle and delicate issue here is that simplicial complex $\mathcal{S}$ (typically but not automatically a finite Constantine Whitehead (CW) complex whose cellular chain complex is endowed with Poincaré duality) is not necessarily a manifold; it is only if the links of all vertices are simplicial spheres, which is, indeed, the very definition of manifold in a piecewise linear context. The difficulty resides in the feature that $n$-spheres are straightforwardly identifiable only for $n = 1, 2$. The problem is tractable for $n = 3$ and possibly 4 only with exponential resources, and it is *undecidable* for $n \geq 5$ [14]. However, given a singular chain complex $\mathcal{S}$, a *normal* map endows it with the homotopy-theoretic global structure of a closed manifold. Sergei P. Novikov proved that for $\dim \mathcal{S} \geq 5$ only the *surgery* obstruction prevents $\mathcal{S}$ from being homotopy equivalent to a closed manifold. The meaning of this is the following: if $\mathcal{S}$ is homotopy equivalent to a manifold then the complex behaves as the base space of a unique Spivak normal fibration, because a manifold has a unique tangent bundle and a unique stable normal bundle. A finite Poincaré complex does not possess such a unique bundle; nevertheless, it possesses an affine fibration that is unique, which is just the Spivak normal fibration. This implies that if $\mathcal{S}$ is homotopy equivalent to a manifold then the spherical fibration associated to the pullback of the normal bundle of that manifold is isomorphic to the Spivak

normal fibration; but the latter has fiber a that is homotopically equivalent to a sphere. This finally entails that all finite simplicial complexes have at least the homotopy type of manifolds with boundary.

We further observe that all available algorithms to compute persistent homology groups are based on the notion of filtered simplicial complex, consisting of pairs: i) the simplex generated at each given step in the recursive construction, and ii) the order-number of the step, a time-like discrete parameter that orders (labels) the collection of complexes by the step at which that simplex appeared in the filtration. The emerging picture can be naturally interpreted as the representation of a *process*, which is endowed with inherent characteristic dynamics that remind us of a discrete-time renormalization group flow [15]. One may then expect that, as it happens with dynamical triangulations of simplicial gravity, the combinatorially different ways in which one may realize the sampling of (*inequivalent*) structures in the persistence construction process, varying the complex shape, give rise to a *natural* probability measure. The measure thus generated is constrained by and must be consistent with the data space invariants and transformation properties.

### 1.5 Step Two: from Data Topology to Data Field

Besides the customary filtrations due to Vietoris-Rips [16], whose $k$-simplices are the unordered $(k+1)$-tuples of points pairwise within distance $\eta$, and to Čech [17], where $k$-simplices are instead unordered $(k+1)$-tuples of points whose $\frac{1}{2}\eta$-ball neighborhoods intersect, or other complexes such as the witness complex [18], which provide natural settings to implement persistence, another filtration, Morse filtration, needs to be considered, that enters here naturally into play.

In the case of those simplicial complexes that are manifold, Morse filtration is a filtration by excursion sets, in terms of what for differentiable manifolds would be curvature-like data. It is indeed a non-smooth, discretized, intrinsic, metric-free version thereof, which is appropriate for the wild simplicial complex that is data space, that can be thought of as the simplicial, combinatorial analogue of the Hodge construction.

It is worth pointing out that, even though it apparently deals with metric-dependent features, in fact Morse filtration is purely topological, namely it is independent on both the Morse function and the pseudo-metric adopted. Also, Morse theory generates a set of inequalities for alternating sums of Betti numbers in terms of corresponding alternating sums of the numbers of critical points of the Morse function for each given index. The analogy with the Hodge scheme is far reaching: simplicial Morse theory generates notions of intrinsic, discrete

gradient vector field and gradient flow, associated to any given Morse function $f_M$. The latter played a particularly significant role – which has been interpreted in the framework of discrete differential calculus – in applications to classical field theory over arbitrary discrete sets [19], which is well described in a non-commutative geometry setting [20].

A Morse complex, built out of the critical points of (any) Morse function with support on the vertex set of $\mathcal{S}$, has the same homology as the underlying structure. This assumes particular importance because the Morse stratification induced [21] is essentially the same as the Harder-Narasimhan [22] stratification of algebraic geometry: one can construct the PL analogue of local *co-ordinates* at the Morse critical points and provide a viable representation of the normal bundle to the critical sets. It helps recalling that the relation between Morse and homology theory is generated by the property that the number of critical points of index $i$ of a given function $f_M$ is equal to the number of $i$ cells in the simplicial complex obtained *climbing* $f_M$, that manifestly bears on $b_i$. Morse homology is isomorphic to the singular homology; Morse and Betti numbers encode the same information, yet Morse numbers allow us to think of the underlying true *manifold*.

Inspired by what happens in the simpler context of gravity, we select the Gromov-Hausdorff (GH) topology [23, 24] to construct a self-consistent measure over $\mathcal{S}$. Gromov's spaces of bounded geometries in fact provide the natural framework to address the measure-theoretical questions posed by simplicial geometry in higher-dimensions. Specifically, it allows us to establish tight entropy estimates that characterize the distribution of combinatorially inequivalent simplicial configurations. In gravity theory the latter problem was solved [25]; however we should keep in mind that one deals with an underlying metric vector space that gives rise, under triangulation, to a simplicial complex, which is a Lorentz manifold.

The GH topology leads naturally to the construction of a statistical field theory of data, as its statistical features are fully determined by the *homotopy* types of the space of data [26]. Complexity and randomness of spaces of bounded geometry can be quite large in the case of Big Data, since the number of *coverings* of a simplicial complex of bounded geometry grows exponentially with the volume. A sort of *thermodynamic limit* then needs to be realized over the more and more random growing filtrations of simplicial complexes. To explain this within the present context, a well defined statistical field theory is required to deal with the extension of the statistical notion of Gibbs field to the case where the substrate is not simply a graph but a simplicial complex, which amounts to proving the property that the substrate underlying the Gibbs field may itself be in some way random. This can be done by resorting to Gibbs *families* [27], so that the ensuing

ensemble of geometric systems – a sort of phase space endowed with a natural measure – behaves as a statistical mechanics object. There ensues the possibility of finding a critical behavior as diversified phase structures emerge – entailing a sort of phase transition when the system passes from one homotopy type to another. The final message is: the deep connection between the simplicial complex structure of data space and the information that such space hides, which is encoded at its deepest levels, resides in the property that data can be partitioned in a variety of equivalence classes and classified by their homotopy type, all elements of each of which encode similar information. In our metaphor, in $X$ information behaves as a sort of *order parameter*.

### 1.6 The Topological Field Theory of Data

A single mathematical object encompasses most of the information about the global topological structure of the data space: the Hilbert-Poincaré series $\mathcal{P}(z)$ (in fact a polynomial in some indeterminate $z$), generating function for the Betti numbers of the related simplicial complex. $\mathcal{P}(z) = \sum_{i \geq 0} b_i z^i$ can be generated through a field theory, as it turns out to be nothing but one of the functors of the theory itself for an appropriate choice of the field action.

The best known analogy to refer to for this formal setup – naturally keeping in mind not only the analogies but mostly the deep structural differences: continuous vs. discrete, tame vs. wild, finite vs. infinite gauge group – is Yang-Mills field theory (YMFT) [28]. In YMFT the variables are a connection field over a manifold $M$ (in this case, a Riemann surface), and the gauge group G is SU(N) (a Lie group of $n \times n$ unitary matrices), under which the Chern-Simons (CS) action (i.e., the $(2k-1)$-form defined in such a way that its exterior derivative equals the trace of the $k$-th power of the curvature) is invariant.

Paraphrasing Terry Tao [29], one may think of a *gauge* as simply a *global coordinate system* that varies depending on one's location over the reference (ambient) space. A gauge transformation is nothing but a change of coordinates *consistently* performed at *each* such location, and a gauge theory is the model for a system whose dynamics is left unchanged if a gauge transformation is performed on it. A global coordinate system is an isomorphism between some geometric or combinatorial objects in a given class and a standard reference object in that same class. Within a gauge-invariant perspective – as all geometric quantities must be converted to the values they assume in that specific representation – it ensues that every geometric statement has to be invariant under coordinate changes. When this can be done, the theory can be cast into a coordinate-free

form. Given the coordinate system and an isomorphism of the standard object, a new coordinate system is simply obtained by composing the global coordinate system and the standard object isomorphism, namely operating with the group of all transformations that leave the gauge invariant. Every coordinate system arises in this manner. The space of coordinate systems can then be fully identified with the isomorphism group $G$ of the standard object. This group is the *gauge group* for the class of objects considered. This very general and simple definition of gauge group allows us to introduce in our scheme a general notion of *coordinates*. These can be straightforwardly identified by the existing intrinsic way to identify mutual relations between objects introduced by the data space topology and by the proximity criterion adopted. It is worth noticing how different such a notion is from the customary notion of coordinates in a vector space.

Let us continue the YMFT analogy. The base-space for YMFT is a smooth manifold, $M$, over which the connection field is well defined and allows for a consistent definition of the action, since the curvature, which is simply the exterior derivative of the connection plus the wedge product of the connection by itself, is well defined everywhere. Field equations in this case are nothing but a *variational machinery* that takes a symmetry constraint as input, expressed as invariance with respect to $G$, and gives as output a field satisfying that constraint. In YMFT, connections allow us to do calculus with the appropriate type of field attaching to each point $p$ of $M$ a vector space – a *fiber* over that point: the field at $p$ is simply an element of such a fiber. The resulting collection of objects (manifold $M$ plus a fiber at every point $p \in M$) is a *vector bundle*. In the presence of a gauge symmetry, every fiber must be a representation (not necessarily different) of the gauge group, $G$. The field structure is that of a $G$-bundle. Atiyah and Bott [30], via an infinite-dimensional Morse theory with the CS action functional as Morse function, in addition to Harder and Narasimhan [22], via a purely combinatorial approach, have both established a formula that expresses the Hilbert-Poincaré series as a functor of the YMFT, in terms of the partition functions corresponding to all Levi subgroups of $G$, a form that is reminiscent of the relation between grand-canonical and canonical partition functions in statistical mechanics.

For the space of data the picture is obviously more complex, because of the more complex underlying structure. Vector bundles of the differential category have a PL category analogue, referred to as *block bundles* [31]. These allow us to reduce geometric and transformation problems characteristic of manifolds to homotopy theory for the groups and the complexes involved. This leads in a natural way to the reconstruction of the $G$-bundle moduli space in a discretized setting. For simplicial complexes that, as already noticed, may not be manifolds, Novikov's lesson is that this can be done just in homotopy terms [14]. Since the homotopy

class of a map fully determines its homology class, the simplicial block-bundle construction furnishes all necessary tools to compute the Poincaré series. Also, in spite of its topological complexity, data space offers a natural, simple choice for the action. Indeed an obvious candidate to start with the exponentiated action is the Heat Kernel $\mathcal{K}$, because the Heat Kernel's trace is precisely proportional to the Poincaré series [32]. $\mathcal{K}$ can be obtained by constructing an intrinsic (metric-free) combinatorial Laplacian over the simplicial complex [33]. This is done by the ad hoc construction of the Hodge decomposition over $\mathcal{S}$ and the related Dirac operator.

An oriented simplicial complex is one in which all simplices in the complex, except for the vertices and empty simplex, are oriented. For any finite simplicial complex $K$ and any nonnegative integer $d$, the collection of $d$-chains of $K$, $\mathcal{C}_d$, is a vector space over $\mathbb{R}$ (nevertheless, the chains still form a group; we refer to the set of chains of a given dimension as the chain group of that dimension). A basis for $\mathcal{C}_d$ is given by the elementary chains associated with the $d$-simplices of $K$, so $\mathcal{C}_d$ has finite dimension $D_d(K)$. If the elements of $\mathcal{C}_d$ are looked at as coordinates relative to this basis of elementary chains, we have the standard inner product on these coordinate vectors, and this basis of elementary chains is orthonormal. The $d$-th *boundary* operator is a linear transformation $\partial_d : \mathcal{C}_d \to \mathcal{C}_{d-1}$.

Each boundary operator $\partial_d : \mathcal{C}_d \to \mathcal{C}_{d-1}$ of $K$ relative to the standard bases for $\mathcal{C}_d$ and $\mathcal{C}_{d-1}$ with some given orderings has a matrix representation $\mathbf{B}_d$. The number of rows in $\mathbf{B}_d$ is the number of $(d-1)$-simplices in $K$, and the number of columns is the number of $d$-simplices. Associated with the boundary operator $\partial_d$ is its adjoint $\partial^*$, $\partial^* : \mathcal{C}_{d-1} \to \mathcal{C}_d$.

It is known that the transpose of the matrix for the $d$-th boundary operator relative to the standard orthonormal basis of elementary chains with the given ordering, $\mathbf{B}_d^t$, is the matrix representation of the $d$-th adjoint boundary operator, $\partial^*$, with respect to this same ordered basis. It is worth recalling that the $d$-th adjoint boundary operator of a finite oriented simplicial complex $K$ is in fact the same as the $d$-th coboundary operator $\delta_d : \mathcal{C}^{d-1}(K,\mathbb{R}) \to \mathcal{C}^d(K,\mathbb{R})$ under the isomorphism $\mathcal{C}^d(K,\mathbb{R}) = \mathrm{Hom}(\mathcal{C}_d(K,\mathbb{R}) \simeq \mathcal{C}_d(K)$.

For $K$ a finite oriented simplicial complex, and $d \geq 0$ an integer, the $d$-th *combinatorial Laplacian* is the linear operator $\Delta_d : \mathcal{C}_d \to \mathcal{C}_d$ given by

$$\Delta_d = \partial_{d+1} \circ \partial_{d+1}^* + \partial_d^* \circ \partial_d. \tag{1.7}$$

As for the group $G$, notice that the space of data has a deep, far-reaching property: it is fully characterized only by its topological properties, neither metric nor geometric, thus – as the objects of the theory have no internal degrees of freedom, they are constrained by the manipulation processes they can be submitted to – there is only one natural symmetry it needs to satisfy, which is the invariance

under all those transformations of data space into itself that do not change its topology and are consistent with the constraints.

This requires a more thorough discussion of homomorphisms of topological spaces. Let $\mathfrak{X}$ be a topological space like the space of data, i.e., a space endowed with some notion of *nearness* between its points. The set $\mathcal{H} = \{\mathfrak{h}\}$ of all homeomorphisms $\mathfrak{h} : \mathfrak{X} \mapsto \mathfrak{X}$ representable as continuous, invertible functions can be thought of itself as a space. $\mathcal{H} = \{\mathfrak{h}\}$ is also a group under functional composition. One can define a topology also on $\mathcal{H}$, space of homeomorphisms $\mathfrak{h}(\mathfrak{X})$. The open sets of $\mathcal{H}$ are made up of sets of functions that map compact subsets $\mathcal{K} \subset \mathfrak{X}$ into open subsets $\mathcal{U} \subset \mathfrak{h}(\mathfrak{X})$ as $\mathcal{K}$ ranges throughout $\mathfrak{X}$, and $\mathcal{U}$ ranges through the images of $\mathfrak{X}$ under all allowed homeomorphisms $\mathfrak{h}$ (completed with their finite intersections – which must be open by definition of topology – and arbitrary unions, that once more must be open). This gives a notion of continuity on the space of functions, so that one can consider continuous deformation of the homeomorphisms themselves: the *homotopies*. The *Mapping Class Group* $\mathfrak{G}_{\mathfrak{MC}}$ is defined by taking *homotopy classes of homeomorphisms*, and inducing the group structure from the functional composition group structure – which is already present on the space of homeomorphisms. This general definition allows us to export the notion of mapping class group to the PL case when $\mathfrak{X}$ is a simplicial complex.

The notion of mapping class group is typically used in the context of manifolds. Indeed, for a given manifold $\mathcal{M}$, $\mathfrak{G}_{\mathfrak{MC}}(\mathcal{M})$ can be interpreted as the group of isotopy classes of automorphisms of $\mathcal{M}$. Thus, if $\mathcal{M}$ is a topological manifold, its mapping class group is the group of isotopy-classes of homeomorphisms of $\mathcal{M}$. In the metric case, if $\mathcal{M}$ is smooth $\mathfrak{G}_{\mathfrak{MC}}(\mathcal{M})$ is the group of isotopy-classes of the *diffeomorphisms* of $\mathcal{M}$. Whenever the group of automorphisms of an object $\mathfrak{X}$ has a natural topology, $\mathcal{M}$, $\mathfrak{G}_{\mathfrak{MC}}(\mathfrak{X})$ is defined as $\mathrm{Aut}(\mathfrak{X})/\mathrm{Aut}_0(\mathfrak{X})$ where $\mathrm{Aut}_0(\mathfrak{X})$ is the *path component* of the identity in $\mathrm{Aut}(\mathfrak{X})$ (in the compact-open topology, path components and isotopy classes coincide); so that there is a short-exact sequence of groups:

$$1 \to \mathrm{Aut}_0(\mathfrak{X}) \to \mathrm{Aut}(\mathfrak{X}) \to \mathfrak{G}_{\mathfrak{MC}}(\mathfrak{X}) \to 1. \tag{1.8}$$

All this implies that the gauge group should be chosen as the semidirect product $\mathcal{G} \wedge \mathfrak{G}_{\mathfrak{MC}}$ of the group $\mathcal{G} \sim \mathcal{P}_{\mathcal{Q}}$ of the *path algebra* associated with the process algebra characteristic of the data set in the representation defined by quiver $\mathcal{Q}$, and the (simplicial analog) $\mathfrak{G}_{\mathfrak{MC}}$ of the *mapping class group* [34] for the space of data.

Recall that *process algebra* refers to the *behavior* of a *system* [35]. A system is indeed anything able to exhibit a behavior, which is the entire collection of events or actions that it can perform, together with the order in which they are executed

and other relevant aspects of this execution, such as timing or probabilities, that define the process. The term algebra refers to the fact that the language used to represent the behavior is algebraic and axiomatic. For this reason operations on processes can be defined in terms of quivers, and their effects can be formally represented in terms of the universal algebra associated with the path algebra.

In analogy with the definition of a group, a *process algebra* is any mathematical structure satisfying the axioms given for its operators. A process is then an element of the universe of the process algebra. The axioms allow calculations with processes. Even if process algebras have their roots in universal algebra, it often goes beyond the bounds of universal algebra: for example, the restriction to a single universe of elements can be relaxed and different types of elements can be used, sometimes, also binding operators. The structure is capable of supporting mathematical reasoning about behavioral equivalences, meaning that whatever the specific approach followed for their definition, these are congruences with respect to behavioral operators.

On the other hand, a process can be modeled as an automaton: an abstract machine with a discrete number of *states* (among which the initial state, not necessarily unique, and the final state) and of *transitions*, i.e., all possible ways of going from a state to its *neighbor* states through the execution of elementary actions, the basic units of a behavior. Then, a generic behavior is an execution path of a number of elementary actions that leads from some initial state to its final state, and an automaton is characterized by the complete set of execution paths. Considering these actions as elements of an alphabet, an automaton is the finite representation of a *formal language*. The important issue of deciding when two automata can be considered equal is in this view expressed by a notion of *semantic equivalence*, specifically of *language equivalence*: two automata are equal when they have the same set of execution paths, or – differently stated – they accept/recognize the same language. In this context, an algebra that allows reasoning about automata is the algebra of regular expressions [36].

Since in automata theory the notion of *interaction* is missing, in order to model a system that interacts with other similar systems, *concurrency theory* is typically used: the theory of interacting, parallel, distributed or reactive systems that provides a process algebra with parallel composition among its basic operators. In this case, the abstract, universal model is the *transition systems* in which the notion of equivalence is not necessarily restricted to language equivalence, but rather to *bisimilarity*. Two transition systems are bisimilar if, and only if, they can mimic each other's behavior in any state they may reach.

Finally, we must take into account that any algebra with a finite number of generators and a finite number of relations can be written as a quiver with relations (though not necessarily in a unique way) by thinking of the set of execution paths

of the automaton's actions as the basis of a *k-path algebra* with composition law induced by the structure of the combinatorial data of a suitable $k$-Quiver ($kQ$). For a given quiver $kQ$, a relation is simply a $k$-linear combination of paths in $kQ$. Given a finite number of relations, one can form their two sided ideal $\mathcal{R}$ in the path algebra, and thus define the algebra $\mathcal{A} \sim kQ/\mathcal{R}$ as a *quiver with relations*. Process algebras can always be assumed to be representable by a quiver with relations. $\mathcal{G}$ is the group associated with $\mathcal{A}$.

A few technicalities are needed here to better define the notion of *process algebra* adopted here. Any finite-dimensional algebra which is *basic* (i.e., all of its irreducible modules are one-dimensional) is isomorphic to a quotient of the path algebra $\mathfrak{P}_{\mathcal{Q}}$ of its quiver $\mathcal{Q}$ modulo an admissible ideal $\mathfrak{I}$.

An analogous, but more subtle, result holds at the basic coalgebra level, through the notion of *path coalgebra* $\mathfrak{C}$ of a *quiver with relations* $(\mathcal{Q}, \mathcal{R})$, where $\mathcal{R}$ is the set of relations induced by $\mathfrak{I}$.

Fix a field, say $\mathcal{K}$. A $\mathcal{K}$-coalgebra – that we shall simply denote as $\mathfrak{C}$ – is a triple $(\mathfrak{C}_{\mathcal{K}}, \Delta, \epsilon)$ consisting of a $\mathcal{K}$-vector space $\mathfrak{C}_{\mathcal{K}}$ and two $\mathcal{K}$-linear maps: the *coproduct* $\Delta : \mathfrak{C}_{\mathcal{K}} \to \mathfrak{C}_{\mathcal{K}} \otimes \mathfrak{C}_{\mathcal{K}}$ and the *co-unit* $\epsilon : \mathfrak{C}_{\mathcal{K}} \to \mathcal{K}$, such that the two equalities hold:

$$(\Delta \otimes \mathbb{I})\,\Delta = \Delta\,(\mathbb{I} \otimes \Delta), \quad (\epsilon \otimes \mathbb{I})\,\Delta = (\mathbb{I} \otimes \epsilon)\,\Delta = \mathbb{I}, \tag{1.9}$$

$\mathbb{I}$ denoting the identity map in $\mathfrak{C}$.

A sub-coalgebra $\mathfrak{A}$ of $\mathfrak{C}$, if it exists, is a $\mathcal{K}$-vector subspace $\mathfrak{A}_{\mathcal{K}}$ of $\mathfrak{C}_{\mathcal{K}}$ such that $\Delta(\mathfrak{A}) \subseteq \mathfrak{A} \otimes \mathfrak{A}$. Henceforth we shall drop index $\mathcal{K}$ whenever it is not necessary.

In this setting *quiver* $\mathcal{Q}$ is a actually a quadruple $(\mathcal{Q}_0, \mathcal{Q}_1, s, e)$, where: $\mathcal{Q}_0$ is a set of vertices and $\mathcal{Q}_1$ a set of *arrows* (oriented edges) in some given ambient space, and for each $\mathfrak{a} \in \mathcal{Q}_1$ the vertices $s(\mathfrak{a})$ and $e(\mathfrak{a})$ in $\mathcal{Q}_0$ are respectively the *source* (start point) and the *sink* (end point) of $\mathfrak{a}$. When $e(\mathfrak{a}) \equiv s(\mathfrak{a})$, arrow $\mathfrak{a}$ is said to be a *loop*.

For $\kappa$ and $\ell$ vertices ($\kappa, \ell \in \mathcal{Q}_0$) an oriented path $\mathfrak{p}_L$ of length $L$ in $\mathcal{Q}$ from $\kappa$ to $\ell$ is the formal ordered composition of arrows

$$\mathfrak{p}_L = \mathfrak{a}_L \circ \mathfrak{a}_{L-1} \circ \cdots \circ \mathfrak{a}_2 \circ \mathfrak{a}_1, \tag{1.10}$$

where $s(\mathfrak{a}_1) \equiv \kappa$, $e(\mathfrak{a}_L) \equiv \ell$, and, for $j = 2, \ldots, L$, $e(\mathfrak{a}_{j-1}) \equiv s(\mathfrak{a}_j)$. Also, to any vertex $\kappa \in \mathcal{Q}_0$ one formally attaches a trivial path of length 0, $\mathfrak{p}_0$, starting and ending at $\kappa$, such that for any arrow $\mathfrak{a} \in \mathcal{Q}_1$ such that $s(\mathfrak{a}) = \kappa$, or $\mathfrak{b} \in \mathcal{Q}_1$ such that $e(\mathfrak{b}) = \kappa$, one has – respectively – $\mathfrak{a} \circ \mathfrak{p}_0 = \mathfrak{a}$, $\mathfrak{p}_0 \circ \mathfrak{b} = \mathfrak{b}$. The set of trivial paths can be identified with the set of vertices $\mathcal{Q}_0$. A path $\mathfrak{c}$ that starts and ends at the same vertex is a *cycle*. Loops are cycles.

Let $\mathcal{H}_{\mathcal{K}\mathcal{Q}}$ be the $\mathcal{K}$-vector space generated by the set of all paths in $\mathcal{Q}$. Endow $\mathcal{H}_{\mathcal{K}\mathcal{Q}}$ with the structure of a $\mathcal{K}$ algebra (note, not necessarily unitary) by defining

the algebra composition law (we may call it *multiplication*) as that induced by simple concatenation of paths: for $\mathfrak{p}_L = \mathfrak{a}_L \circ \cdots \circ \mathfrak{a}_1$, $\mathfrak{q}_M = \mathfrak{b}_M \circ \cdots \circ \mathfrak{b}_1$,

$$\mathfrak{p}_L \bullet \mathfrak{q}_M \doteq \begin{cases} \mathfrak{a}_L \circ \cdots \circ \mathfrak{a}_1 \circ \mathfrak{b}_M \circ \cdots \circ \mathfrak{b}_1, & \text{if } e(\mathfrak{b}_M) \equiv s(\mathfrak{a}_1), \\ \emptyset, & \text{otherwise.} \end{cases} \tag{1.11}$$

The algebra $\mathfrak{P}_{\mathcal{KQ}}$ thus generated is the *path algebra* of the quiver $\mathcal{Q}$.

$\mathfrak{P}_{\mathcal{KQ}}$ has a natural *grading*:

$$\mathfrak{P}_{\mathcal{Q}} \equiv \mathfrak{P}_{\mathcal{KQ}} = \mathfrak{P}_{\mathcal{Q}_0} \oplus \mathfrak{P}_{\mathcal{Q}_1} \oplus \cdots \oplus \mathfrak{P}_{\mathcal{Q}_m} \oplus \cdots, \tag{1.12}$$

where $\mathcal{Q}_m$ denotes the set of all paths of length $m$, $\mathcal{Q}_m = \{\mathfrak{p}_m \,|\, m \in \mathbb{N}\}$, that form a complete set of primitive, orthogonal idempotents of $\mathfrak{P}_{\mathcal{Q}}$.

$\mathfrak{P}_{\mathcal{Q}}$ is unitary if $\mathcal{Q}_0$ is finite; $\mathfrak{P}_{\mathcal{Q}}$ is finite-dimensional if and only if $\mathcal{Q}$ is finite and has no cycles.

An ideal $\mathfrak{I} \subseteq \mathfrak{P}_{\mathcal{Q}}$ is called *ideal of relations* if $\mathfrak{I} \subseteq \mathfrak{P}_{\mathcal{Q}_2} \oplus \mathfrak{P}_{\mathcal{Q}_3} \oplus \cdots \doteq \mathfrak{P}_{\mathcal{Q}_{\geq 2}}$.

For $\mathcal{Q}$ finite, ideal $\mathfrak{I}$ of $\mathfrak{P}_{\mathcal{Q}}$ is *admissible* if and only if there exists an integer $n \geq 2$ such that, denoting by $\mathfrak{P}_{\mathcal{Q}_{\geq n}}$ the ideal $\mathfrak{P}_{\mathcal{Q}_{\geq n}} \doteq \mathfrak{P}_{\mathcal{Q}_n} \oplus \mathfrak{P}_{\mathcal{Q}_{n+1}} \oplus \cdots$, one has $\mathfrak{P}_{\mathcal{Q}_{\geq n}} \subseteq \mathfrak{I} \subseteq \mathfrak{P}_{\mathcal{Q}_{\geq 2}}$.

Finally, a *quiver with relations* $\mathcal{Q}_{\mathcal{R}}$ is a pair $(\mathcal{Q}, \mathcal{R})$, namely a quiver $\mathcal{Q}$ endowed with the ideal generated by the relations $\mathcal{R}$ induced by $\mathfrak{I}$. If $\mathfrak{I}$ is admissible then $\mathcal{Q}_{\mathcal{R}}$ is a *bound quiver*.

If for $\mathfrak{p}_L = \mathfrak{a}_L \circ \mathfrak{a}_{L-1} \circ \cdots \circ \mathfrak{a}_2 \circ \mathfrak{a}_1$ a path of length $L$ in $\mathcal{Q}$ from vertex $\kappa$ to vertex $\ell$ one defines:

$$\begin{aligned} \Delta(\mathfrak{p}_L) &\doteq \mathfrak{p}_0^{(\ell)} \otimes \mathfrak{p}_L + \mathfrak{p}_L \otimes \mathfrak{p}_0^{(\kappa)} + \sum_{j=1}^{L-1} \mathfrak{a}_L \circ \cdots \circ \mathfrak{a}_{j+1} \otimes \mathfrak{a}_j \circ \cdots \circ \mathfrak{a}_1 \\ &\doteq \sum_{\substack{\mathfrak{r},\,\mathfrak{s} \\ \mathfrak{r} \bullet \mathfrak{s} = \mathfrak{p}_L}} \mathfrak{r} \otimes \mathfrak{s}, \end{aligned} \tag{1.13}$$

whereas, for any trivial path $\mathfrak{p}_0$, $\Delta(\mathfrak{p}_0) = \mathfrak{p}_0 \otimes \mathfrak{p}_0$; and

$$\epsilon(\mathfrak{p}) \doteq \begin{cases} 1, & \text{if } \mathfrak{p} \in \mathcal{Q}_0, \\ 0, & \text{if } \mathfrak{p} \text{ is a path of length } \geq 1, \end{cases} \tag{1.14}$$

then $(\mathfrak{P}_{\mathcal{KQ}}, \Delta, \epsilon)$ is the path coalgebra of quiver $\mathcal{Q}$ (or $\mathcal{Q}_{\mathcal{R}}$ if $\mathcal{Q}$ is endowed with relations $\mathcal{R}$). This completes the toolkit necessary for the construction of the factor $\mathcal{G} \sim \mathcal{P}_{\mathcal{Q}}$ of the gauge group.

As for $\mathfrak{G}_{\mathfrak{MC}}$, a few extra comments are needed to clarify how one has to proceed to coherently and practically construct its simplicial complex representation. A

key aspect here is the set of actions of the mapping class groups on spaces of different sorts, encoding characteristic geometric and topological features. Among these homotopy classes, foliations, conformal structures have all been extensively studied [37]. All these actions are induced by corresponding actions of the homeomorphisms of the base space on the objects selected. Moreover, the spaces on which the mapping class groups act can be equipped with different structures, e.g., groups, simplicial complexes, or manifolds, and the mapping class groups are embedded accordingly into groups of algebraic isomorphisms, simplicial automorphisms, isometries of the related metrics – if any. For most of these actions, the natural homomorphism from the mapping class group to the automorphism group of the given structure is an isomorphism. Among these, particularly interesting in the present context are the actions by simplicial automorphisms on the abstract simplicial complexes associated to $X$; namely, actions by piecewise linear automorphisms of the associated measured foliations space, equipped, for example, with the train-track piecewise linear structure introduced by Thurston [38] or with the set of self-preserving intersection functions.

The latter structure is related with the braid group, whose central extension – not unexpectedly – is the group of permutations. $\mathfrak{G}_{\mathfrak{MC}}$ is finite and finitely presented; its presentation, as well as its representations, can be completely constructed once one knows the full homotopy of the simplicial complex. Recently, a complete representation of $\mathfrak{G}_{\mathfrak{MC}}$ realized in terms of the group $SU(1,1)$ of hyperbolic rotations has been obtained by the authors (and is reported in [39]).

We claim that, in spite of the formal difficulties, mimicking the block bundle approach for the appropriate simplicial complex structure and given $G$, the data space topological invariants (among which Betti numbers) can be computed in the context of the proposed field theory through the (recursively computable) subsets of symmetries of $\mathcal{G} \wedge \mathfrak{G}_{\mathfrak{MC}}$. The benefit is twofold. On the one hand the cosets of $\mathcal{G} \wedge \mathfrak{G}_{\mathfrak{MC}}$ order data in equivalence classes with respect to isotopy, leading to a canonical system in the related process algebras. On the other hand, one can make a unique choice among the several possible theories – the multiplicity being related with the plurality of topological structures due to the passage through Morse numbers (Morse and Betti numbers are related through inequalities, not equalities) – in the following way. One begins by constructing, for all manifolds in the family generated by the collection of Morse numbers, the *free* field theories whose exponentiated action is simply the Heat Kernel, for which the partition function is the generating function of the manifold Betti numbers. By self-consistency, i.e., simply comparing the coefficients of $\mathcal{P}(z)$ with the Betti numbers outcome of the *phenomenological* persistent homology one identifies which is the effective data manifold.

In this way, not only do we fully recover through the construction proposed the whole data space topology (for example, the set of Betti numbers), but we are able to continue to construct an autonomous, self-consistent topological data field theory (TDFT) on the space of data: once more *the fascination of unexpected links in mathematics* [40].

As a final remark, notice that the resulting picture comprises a surprising amount of information on the associated moduli spaces as well; markedly, the quiver representation for the path algebra $\mathcal{A}$ (see also [41, 42]), basic tools for the description of processes involving maps and transformations of data sets.

## 1.7 The Formal Language Theory Facet

The construction outlined so far naturally brings to light a new facet: formal language theory, which conveys a dimension as much unexpected as elegant in its form.

A preliminary question to raise at this point is whether the adopted topological landscape is inherently coherent with the structure of Formal Language Theory (FLT). As we know, a central issue in the theory of computation is to determine classes of languages whose representation has finite specification [36]. A formal language defined over a finite alphabet $\mathfrak{A}$ of symbols is a subset of the set $\mathfrak{A}^*$ of all strings of any length that can be represented by that alphabet. As a consequence, the number of possible representations is countably infinite and the set of all possible languages over a given alphabet $\mathfrak{A}$ is uncountably infinite. Under these conditions we are obviously unable to represent all languages. Coupled with this issue there is the limit posed by well-known Gold's theorem for which the *minimum automaton identification from given data is NP-Complete* [43]. In the TDFT context, the challenge is to construct a finite representation of the language defined over the alphabet whose symbols are the generators of gauge group $G$, and whose cosets partition the data space $X$ in equivalence classes of finitely presented objects. Such languages can be finite or infinite; what is interesting here is that their presentation can always be finitely given in $\mathcal{G} \wedge \mathfrak{S}_{\mathfrak{MC}}$. In other words, such languages are each a collection of discrete spaces containing a finite number of homeomorphic objects; by the TDFT we construct a language of data, the language proper to topological shape $\mathcal{S}$.

Interpreting the gauge group $G$ as topological shape language requires a resort to a notion of duality somehow similar to that entering the construction of Langland's dual group, yet designed to represent the relationship between structure and function of a behavior: a *mirror* symmetry that allows each to affect the other

in the same way. As a consequence, we can characterize the data language as the process algebra whose processes are well-behaved with respect to *modulo bisimulation* [44], by attributing them the same, unique (bi-)algebra induced by the gauge group $\mathcal{G} \wedge \mathfrak{G}_{\mathfrak{MC}}$, with $\mathcal{G}$, as mentioned, the group of $\mathcal{A}$.

The role of $\mathfrak{G}_{\mathfrak{MC}}$ in the discrete case can be naturally traced back to *Automatic Groups* [45], i.e., finitely generated groups equipped with several finite-state automata that are able to distinguish whether or not a given word – a representation of a group element – is in *canonical form*, and hence if two elements in canonical form differ, and if they do, by which generators. It may be worth recalling that automatic groups were originally introduced in connection with topology, in particular with the study of the fundamental group, and of the homotopy (3-manifolds), because the class of automatic groups can be extended to include the fundamental group of every compact 3-manifold, thus satisfying Thurston's geometrization [38]. In the topological structure we are dealing with here – where we consider collections of *relational* simplexes, built by combinatorially gluing together relational simplices – the task is much more complex. However, as the basic structure is fully controlled by homotopy types, turning the generation of a family of parametrized simplicial complexes into a classification problem in FLT is natural and straightforward in its statement, if not in its solution. One should be aware, however, that issues of uncontrollable algorithmic complexity or even of undecidability may possibly arise.

Moreover, the syntax of a language in FLT is traditionally described by using the notion of grammar, defined by the relations that are necessary to build correct syntax constructs from atomic entities (symbols). This is what allows us to describe the syntax of a formal language universally, in spite of the representation of its texts. In addition, the syntax constructs are typically described as resorting to the notion of syntax diagram, $\mathbb{D}$, that is the connected multigraph with nodes labeled in terms of the formal language's alphabet $\mathfrak{A}$ and connections – in our representation not only edges or links, but also higher-dimensional simplices – that represent the syntax relations. The multigraph of a syntax diagram may be directed or not, and in view of its combinatorial structure, inherited from the simplicial structure of data space and accounted for in the FTL vision, it is itself to all effects a simplicial complex. It is possible to select specific syntax diagrams (referred to as *correct*, as defined below) out of the set of all syntax diagrams on $\mathfrak{A}$ to construct different grammars. The formalism used to do this requires a fundamental notion: that of neighbor grammars [46], whose meaning is the following. Define for each $\mathbb{D}$, the collection of subdiagrams labeled by the set of pairs $(\mathbb{D}', \mathfrak{s})$ where $\mathbb{D}' \subseteq \mathbb{D}$ is another syntax diagram and $\mathfrak{s}$ is the inclusion map of $\mathbb{D}'$ into $\mathbb{D}$. The neighborhood of a symbol of $\mathfrak{A}$ is a syntax diagram that contains

the node singled out by this symbol. The neighbor grammar of the given grammar consists of the finite family of neighborhoods defined for each symbol of $\mathfrak{A}$. A given syntax diagram is said to be *correct* if for each of its nodes, labeled by some symbol of $\mathfrak{A}$, it includes some a neighborhood of this symbol. Such a neighborhood should contain all simplices adjoining to its center. There is therefore at least one cover consisting of neighborhoods for each correct syntax diagram in the given neighbor grammar. Such cover is the *syntax*. Furthermore, the category $\mathfrak{D}$ of syntax diagrams over the given alphabet can be introduced, based on the neighboring grammar. It is known [46] that the category of correct syntax diagrams, defined as $\mathfrak{D}$ but limited to correct syntax diagrams, admits a Grothendieck topology [47].

It is the formal language generated by the field theory through its gauge group that makes the TDFT consistent with a formal language architecture. This comes exactly from the property of having a Grothendieck topology at our disposal. Indeed, the Grothendieck topology is a structure on a category $\mathcal{C}$ which makes the objects of $\mathcal{C}$ behave like the open sets of a topological space $\mathcal{X}$. Its characteristic is that it replaces the notion of a collection of open subsets of $\mathcal{U} \subseteq \mathcal{X}$ which is stable under inclusion by the notion of a *sieve*. If $c$ is an object in $\mathcal{C}$, a sieve $\mathfrak{S}$ on $c$ is a *subfunctor* of the functor $\mathrm{Hom}(-,c)$ – i.e., for all objects $c \in \mathcal{C}$, $\mathfrak{S}(c) \subseteq \mathrm{Hom}(c,c)$, and for all arrows $f : c \to c$, $\mathfrak{S}(f)$ is the restriction of $\mathrm{Hom}(f,c)$, *pullback* by $f$ to $\mathrm{Hom}(c,c)(c)$: the Yoneda embedding applied to $c$. In the case of $\mathcal{O}(\mathcal{X})$ [the category whose objects are the open subsets $\mathcal{U} \subseteq \mathcal{X}$ and whose morphisms are the inclusion maps $\mathcal{V} \to \mathcal{U}$ of open sets $\mathcal{U}$ and $\mathcal{V}$ of $\mathcal{X}$], a sieve $\mathfrak{S}$ on an open set $\mathcal{U}$ just selects the collection of open subsets of $\mathcal{U}$ which is stable under inclusion. If $\mathcal{W} \subset \mathcal{V}$, then there is a morphism $\mathfrak{S}(\mathcal{V}) \to \mathfrak{S}(\mathcal{W})$ given by composition with the inclusion $\mathcal{W} \to \mathcal{V}$. If $\mathfrak{S}(\mathcal{V})$ is non-empty, there follows that $\mathfrak{S}(\mathcal{W})$ is also non-empty. The pullback of $\mathfrak{S}$ along $f$, that we denote by $f * \mathfrak{S}$, is – for $\mathfrak{S}$ a sieve on $\mathcal{X}$ and $f : \mathcal{Y} \to \mathcal{X}$ a morphism, left composition by $f$ – the sieve on $\mathcal{Y}$ defined as the fibered product $\mathfrak{S} \times_{\mathrm{Hom}(-,\mathcal{X})} \mathrm{Hom}(-,\mathcal{Y})$ together with its natural embedding in $\mathrm{Hom}(-,\mathcal{Y})$. More concretely, for each object $\mathcal{Z}$ of $\mathcal{C}$, $f * \mathfrak{S}(\mathcal{Z}) = \{g : \mathcal{Z} \to \mathcal{Y} \,|\, f\,g \in \mathfrak{S}(\mathcal{Z})\}$, and $f * \mathfrak{S}$ inherits its action on morphisms by being a subfunctor of $\mathrm{Hom}(-,\mathcal{Y})$. Finally, a Grothendieck topology $\mathfrak{G}_{\mathcal{C}}$ on a category $\mathcal{C}$ is a collection, for each object $c \in \mathcal{C}$, of distinguished sieves on $c$, say $\mathfrak{G}_{\mathcal{C}}(c)$, called covering sieves of $c$. The selection process, whereby such collection is generated, will be subjected to a number of axioms. A sieve $\mathfrak{S}$ on an open set $\mathcal{U} \in \mathcal{O}(\mathcal{X})$ will be a covering sieve if, and only if, the union of all the open sets $\mathcal{V}$ for which $\mathfrak{S}(\mathcal{V})$ is non-empty and equals $\mathcal{U}$; in other words, if and only if $\mathfrak{S}$ gives us a collection of open sets which cover $\mathcal{U}$ in the customary sense.

### 1.8 Language, Structure and Behavior, Automata

The TDFT construct has crucial consequences in terms of theoretical computer science. In particular, the three basic identifications it implies have a far reaching interpretation: i) the *architectural structure* of the dataset seen as a *G-fiber bundle*, consisting of a base space, the space of data $X$, dealt with as a topological space, a fiber attached to each point of $X$, the set of *fibers*, each as a representation of the *gauge group* $G = \mathcal{G} \wedge \mathfrak{G}_{\mathfrak{Mc}}$; ii) the *field* as an element of the *fiber* at each point of the data space; iii) an *action* – in the simplest non-interacting case is the combinatorial Laplacian – able to describe the processes over data as transformations of the global topological landscape.

This architecture is indeed what allows us to touch the final goal: the definition of a universal methodology whereby, starting from the exploration of (large) data sets, we may construct a *language* capable of describing processes over data as a unified operational system of structure and behavior. This new object can be interpreted as *true* (effective, extended) *data space*, which includes, besides the topological features inherent in the data set, the set of all possible transformations allowed on data, which are generated by the group of all its possible topology-preserving transformations as well as by the related process algebra and reflected in the resulting equivalence classes. In such perspective, the system becomes itself a *self-organizing program*, whose identifiers are the interactions that characterize the field action. Such interactions correlate parts of potential processes (embedded programs) of real life applications: a feature typically caught in the $S[B]$ paradigm [48].

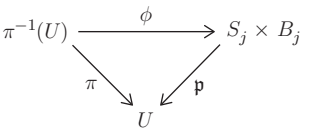
The principle of *self-organization* has long entered as a fundamental feature in the theory of nonlinear, possibly discrete, dynamical systems. It provides the clue to obtain diverse representations of the relation between lower-level elements and higher-order structures in (multi-level) complex systems. Its basic idea is that the interactions among low-level elements, in which each element adjusts to the others, is local because it does not make reference to patterns that are global. It is however this latter feature that leads to the emergence of highly coherent structures and complex behavior over the system as a whole. Such structures, in turn, are able to provide correlations for the lower-level elements with no need of higher-order agents to induce their emergence [50, 51]. In other words, rather than being imposed from above or from outside, the higher-order structures emerge from the interactions internal to the system or between the system and its environment.

From an algebraic perspective, it is the language signature that becomes a measure of the interactions which generate the environment associated with the data set. This is exactly what happens in the $S[B]$ model when one establishes which states connected to $B$ (i.e., which *behavior*) satisfy the constraints imposed

by the set of states $S$ (i.e., the states defining the system's *structure*). In TFTD this is equivalent to the process of determining which is the global fiber bundle to which a given element in a fiber attached to a specific point of the topological space belongs.

In this perspective, the structure of $S[B]$ can be naturally identified as a fiber bundle: $S[B] = (B, S, \pi, \mathcal{B} \doteq \{B_j \,|\, j \in \mathcal{J}\})$, with total space $B$, base space $S \sim X$ (the different notation $S$ is to remind us the we are dealing with a simplicial complex), projection map $\pi : B \to S$ and fiber set, $\mathcal{B}$. $\mathcal{J}$ is a label set tagging points $x_j \in S, j \in \mathcal{J}$. In $\mathcal{B}$ each single fiber $B_j$ specifies the global topological constraints conditioning all the correlations of the $x_j$. It should be recalled here that $S$ is a higher-dimensional *standard* object that provides the frame for the data space $X$. This defines the internal homeomorphisms within the equivalence classes on the fiber, for any subset of constraints corresponding to a given choice of the global invariants.

Fiber $B_j$ is the topological space of *computations* induced by those constraints in $S$ compatible with the fiber structure, whose subset $S_j = \pi^{-1}(B_j)$ is itself a subspace of $S$. It is the commutativity of diagram $\mathfrak{D}_j$:

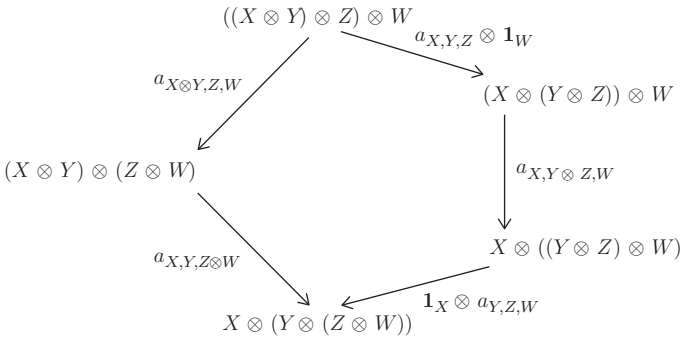



that allows us to identify the homeomorphism $\mathfrak{p}$ as the projection map that establishes the one-to-one relationship restricting $S_j$ to the subfiber $U$ of $S[B]$ that we can finally denote as $S_j[B_j]$. The latter has the same topological invariants as $S_j \leftrightarrow x_j$, so that $\mathfrak{p}$ actually *entangles* computation and its context (i.e., the objects living in $S_j \times B_j, \forall j \in \mathcal{J}$).

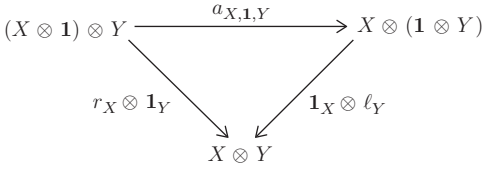
Our TDFT can in this way be viewed as generated by symmetric monoidal functors from the monoidal pseudo $n$-fold category to a monoidal $n$-fold category of spans of sets. The possible resulting degeneracy (more than a single automaton associated with the same language; i.e., strongly connected oriented graphs) reflects the non-uniqueness at the simplicial complex level of the correspondence *Betti numbers to Morse numbers* at the field theoretical level. In the present scheme it is resolved by self-consistency.

Before proceeding, let us recall a few definitions. First, a *Tensor Category* (TC) is a sextuple $(\mathcal{C}; \otimes; a_{\dots}; \mathbf{1}; \ell_{\cdot}; r_{\cdot})$, where $\mathcal{C}$ is a category; operation $\otimes : \mathcal{C} \times \mathcal{C} \longrightarrow \mathcal{C}$ is a (bi)functor; $a_{XYZ} : (X \otimes Y) \otimes Z \simeq X \otimes (Y \otimes Z)$ is a (functorial) associativity constraint; $\mathbf{1}$ is the unit object; while $\ell_X : \mathbf{1} \otimes X \simeq X$ and $r_X : X \otimes \mathbf{1} \simeq X$, subject to a number of axioms. Considering only $\mathbb{C}$-linear abelian tensor categories (with bilinear tensor product), the TC must satisfty two sets of basic (defining) axioms:

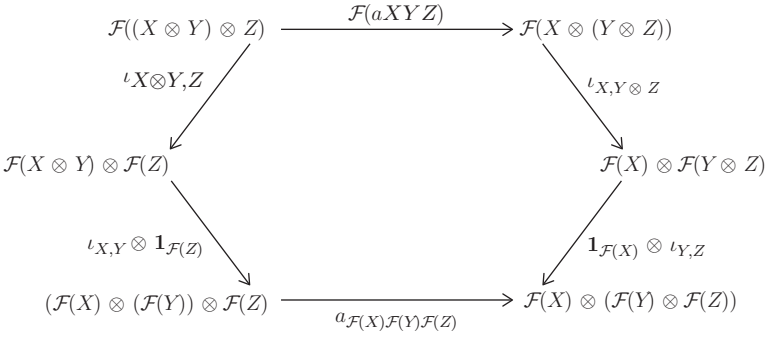
the pentagon axiom:



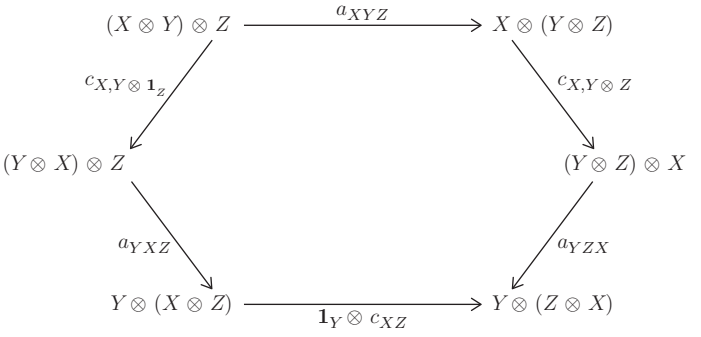

and the triangle axiom:




Two categories $\mathcal{C}_1, \mathcal{C}_2$ are said to be *tensor equivalent* if there exists a functor $\mathcal{F}: \mathcal{C}_1 \longrightarrow \mathcal{C}_2$, together with an isomorphism $\mathcal{F}(\mathbf{1}) \simeq \mathbf{1}$ and a functorial isomorphism $\iota_{X,Y}: \mathcal{F}(X \otimes Y) \longrightarrow \mathcal{F}(X) \otimes \mathcal{F}(Y)$, such that



If moreover a functional isomorphism $c_{XY} : X \otimes Y \simeq Y \otimes X$ exists, satisfying – both for $c_{XY}$ and $c^{(rev)}_{XY} \equiv c^{-1}_{YX}$ – the hexagon axiom:




then $\mathcal{C}$ is a called a *braided tensor category*, and the pure braid group $\mathcal{PB}_n$ acts on $X_1 \otimes X_n$, whereas the braid group $\mathcal{B}_n$ acts on $X^{\otimes n}$. A braided tensor category is symmetric if $c_{XY} \circ c_{YX} = id$ (i.e., $c^{(rev)} \equiv c$), $\forall X, Y$.

For $\mathcal{C}$ a TC, a *module category* over $\mathcal{C}$ is a quadruple $(\mathcal{K}, \otimes, a_{\dots}, \ell_{\cdot})$, $\mathcal{K}$ being a $\mathbb{C}$-linear category and the (exact) bifunctor $\otimes$ denoting now the operation $\otimes : \mathcal{C} \times \mathcal{K} \longrightarrow \mathcal{K}$, satisfying the pentagon and triangle axioms.

An important example of this construction comes from conformal field theory (CFT). In statistical field theory a *conformal theory* is fully determined by its correlation functions, exactly like it happens in TDFT and in $S[B]$. In CFT, correlation functions are bilinear combinations of conformal blocks (sets of correlators that implement the identities and constraints that follow from the global gauge symmetries of the theory), and a monodromy for conformal blocks arises that is encoded into a modular tensor category $\mathfrak{T}$. Given conformal blocks with monodromy described by $\mathfrak{T}$, specifying the correlation functions is equivalent to selecting another category, the *module category* $\mathfrak{M}$ over $\mathfrak{T}$. Also, in a CFT conformal blocks are controlled by a *vertex* algebra $\mathcal{V}$ [52]. A deep theorem [53] states that for $\mathfrak{M}$ to be indecomposable over the representations of $\mathcal{V}$ one can combine conformal blocks of $\mathcal{V}$ into a globally consistent system of correlation functions.

In this complex construction a crucial notion emerges: that of the (asynchronously) $\mathfrak{L}$-combable group [54]. The latter is a group to each element of which we can associate a word in some free group within an arbitrary, abstract family of languages $\mathfrak{L}$. The nature of $\mathfrak{L}$ is rather flexible: it can be the family of regular languages, context-free languages, or indexed languages. Words representing group elements in some of these languages [55] describe (flow-like) transformations over the data set. The class of combable regular languages consists of precisely those groups that are asynchronously automatic. Recalling

the Atiyah-Bott and Harder-Narashiman results for manifolds, it is relevant to try and classify the (normal) sub-groups of $G$, and this can be done in the group and language theoretical setting.

In the algebraic theory of languages, a regular language is fully represented by its syntactic monoid (meaning that the properties of that language, e.g., the expressive power of its first-order logic, are fully contained in the structure of the monoid), which is typically finite. In this framework regular languages are referred to as *languages of data words*. A rigorous, but simple construction, of data words consists in identifying first the alphabet, say $\mathfrak{A}$, and focusing then the attention on words and languages over $\mathfrak{A}$ and on the algebraic theory they generate. The field theoretical construction of the Betti number generating function for data sets is an instance of representation of the complex language of data words associated with the simplicial realization of $\mathfrak{G}_{\mathfrak{MC}}$, which is known to be combable (though in some cases possibly not automatic) [56].

Automata models can be developed for languages of data words, whose basic feature is that they provide a trade-off among three crucial properties: strong *expressivity*, good *closure* properties and decidable (or efficiently decidable) *emptiness*. It strikes an acceptable balance in the trade-off. Logics have been developed to establish the properties of data words: in particular a language of data words is definable in first-order logic *if* its syntactic monoid is aperiodic; a statement that links the feature of definability in first-order logic to a property that in our framework is dynamical.

In the topological setting, the relevant emerging relationships naturally involve invariants that are related not only to objects but maps between pairs of objects as well. Once again we find here an explicit manifestation of *functoriality*. This is the way in which the theories of automata and formal languages merge with the field theoretical picture, because the field theory generates sequences of symbols that enter into play in the simplicial construction of the $G$-bundle associated with the gauge group $G$, as well as the relations among them. In turn, this bears on the enumerative combinatorics content of the theory (because $G$ is reduced essentially to homotopy braids) that provides the language recognized by the automata. Also, combinatorics on words pertains to the wide set of natural operations on languages, in particular to the property – crucial for the final step of pattern discovery in data space – that the orbit of any language in $\mathfrak{L}$ under the *monoid* generated by such a set is finite and bounded, independently of what $\mathfrak{L}$ is.

The use of formal languages leads to the recognition of automatically generated domain-specific languages. The latter are languages appropriate to single out specific topological objects (*concepts*) and their mutual relations, hidden in the noisy landscape of the large data space, and to manage, query and reason over those concepts so as to infer new knowledge. This recalls Codd's theory of database

management with its basic tool, *relational algebra* – derived from the algebra of sets and first-order logic when dealing with finite relations closed under specific operations. Codd's approach tackles the problem top-down, first defining the conceptual model, then classifying data through relations, and finally manipulating such relations through their schemas. The approach based on the topology of data space, on the contrary, tackles the problem bottom-up. The two approaches can thus be associated to two different, complementary ways of thinking: the former, based on the assumption that the agent knows a priori, at least in part, the properties of data (characteristic, e.g., of artificial intelligence approaches to data mining, such as *machine learning*); the latter aimed at inferring new knowledge for the agent, extracting from data (*ontological emergence*) those relations that define hidden structural knowledge-generating patterns, but with no a priori information on what data is about.

The dialectical question about the nature of *patterns*, grounded in the antithesis between pattern *recognition* and pattern *discovery*, has guided us naturally – in the field theoretical context – to search for a way to describe patterns at the same time algebraic, computational, intrinsically probabilistic, yet causal. In TDFT, patterns can be collected in ensembles resorting to equivalence classes of histories, or of sets of states. The strength of such patterns (e.g., their predictive, i.e., information retrieval, capability) and their statistical complexity (via state entropy, or the amount of information retained) provide, for each particular process, a measure of the forecasting ability of the theory over the entire data space.

## 1.9 Emergence of Patterns

We need now to finally merge all the above ingredients into a unique field-theoretical picture, consistent not only with the representation of the space of data equivariant with respect to the transformation properties induced by the simplicial topological scheme itself and by the processes the system may undergo, but also on the full set of characteristic patterns within the data set – via the field correlation functions. The weights depend on the notion of proximity adopted, on the formal language on which the theory is based, on the field action functional selected and on the Morse stratification corresponding to it, as well as on the set of transformations of the data space into itself that preserve its topology. The choice of correlations to represent patterns is crucial: it enables us to make predictions without violating the unavoidable restriction (a mixture of the second law of thermodynamics with the principle of relativity) that predictions can only be based on the process's past, not on any outside source of information except the data

in $X$. In a perspective of this sort, patterns belong to the intrinsic structure of the process, not to the rest of the universe; aggregated pieces of information that share a common structure, and say little about what that pattern is. This is just what correlations are about.

Patterns as represented by field correlations are: *robust*, because they are derived from persistent homology (mediated, if necessary, by the statistical mechanics manipulation process, e.g., smoothing out the role of very high order topological invariants) and hence free, to any desired accuracy, of irrelevant noisy components; *global*, as they describe deep lying correlations dictated by the non-local features of the space topology inherited by the field; *optimal*, based as they are on the variational principle proper to the field theory; *flexible*, due to the vast diversity inherent in their language theoretic structure. This is why they provide essential strategic directions as how to search data space. Whilst several details of the theory remain to be exhaustively worked out, its grand design does not. Of course several of its subtle technicalities need to be completed. A number of applications have started to confirm its potential reach and validity. Among these we mention in particular two: the formulation of a novel *many body* approach to the construction of an effective immune system model [48], and the analysis of the nature of altered consciousness in the *psychoactive drug controlled state* based on functional magnetic resonance imaging data [7, 57].

## 1.10 Conclusions

To conclude, we have outlined the construction of a topological gauge field theory for data space when these data encode information. Such a theory is capable of acting as a machine whose inputs are a space of data and the symmetry group generated by its simplicial complex approximation as resulting from persistent homology, while its output consists of sets of patterns in the form of field correlations as generated by the field equations. These correlation functions fully encode information about patterns in data space, where the relevant information about the system which the data refer to is encoded. The field theory is self-consistent. It is topological because the data space features it resorts to are topological invariants, and because the gauge group embodies the most general transformations of data space, which leave such global topological features unchanged. Finally, the field evolution – due to the PL nature of the construct – has a natural implementation in terms of finite state automata, which maps both the emergence of patterns and the identification of correlations into well-defined formal language theoretical questions.

## 1.11 Acknowledgments

The financial support for this paper was provided by the Future and Emerging Technologies (FET) program within the Seventh Framework Programme (FP7) for Research of the European Commission, under the FET-Proactive grant agreement TOPDRIM, number FP7-ICT-318121.

## References

[1] V. Cerf, *Where is the science in computer science?* Communications of the ACM. 5 (10), 5 (2012).
[2] Various Authors, *Special Section: Dealing with Data*, Science. 331 (2011).
[3] S. Barry Cooper, *Incomputability after Alan Turing*, Notices, AMS. 59(6), 776–784 (2012).
[4] J. Pearl, *Causality: models, reasoning, and inference*. Cambridge: Cambridge University Press, (2009).
[5] L. Wittgenstein, *Philosophical investigations*, transl. Anscombe, GEM. London: Blackwell Publishing, (1921).
[6] J. Hopcroft and R. Kannan, *Foundations of data science*. (2013). Available at http://blogs.siam.org/the-future-of-computer-science/.
[7] G. Petri, M. Scolamiero, I. Donato and F. Vaccarino, *Topological strata of weighted complex networks*, PLoS ONE. 8(6), (2013). e66506. DOI: 10.1371/journal.pone.0066506.
[8] G. Carlsson, *Topology and data*, Bulletin of the AMS. 46(2), 255–308 (2009).
[9] H. Edelsbrunner and J. Harer, *Computational topology: an introduction*. American Mathematical Society, (2010).
[10] A. J. Zomorodian, *Topology of computing*. Cambridge: Cambridge University Press, (2009).
[11] S. Basu, R. Pollack and M. F. Roy, *Algorithms in real algebraic geometry*. New York: Springer-Verlag, (2006).
[12] E. Witten, *Quantum field theory and the Jones polynomial*, Communications in Mathematical Physics, 121(3), 351–399 (1989).
[13] C. R. Shalizi and J. P. Crutchfield, *Computational mechanics: pattern and prediction, structure and simplicity*, Journal of Statistical Physics, 104(3–4), 816–879 (2001).
[14] S. P. Novikov, *On manifolds with free abelian fundamental group and applications*, Izv. Akad. Nauk SSSR ser. mat. 30(1), 208–246 (1966). English translation: A.M.S. Transl. 67(2) 1–42 (1967).
[15] J. Zinn-Justin, *Quantum field theory and critical phenomena*. Oxford: Clarendon Press, (2002).
[16] L. Vietoris, *Über den höeren Zusammenhang kompakter Räume und eine Klasse von zusammenhangstreuen Abbildungen*, Math. Ann. 97, 454–472 (1927).
[17] E. Čech, *Théorie générale de l'homologie dans un espace quelconque*, Fund. Math. 19, 149–183 (1932).
[18] V. de Silva, *A weak definition of Delaunay triangulation*. (2003). *Preprint* arXiv cs. CG/031003 v1.
[19] B. Auchmann and S. Kurz, *A geometrically defined discrete Hodge operator on simplicial cells*, IEEE Transactions on Magnetics. 42(4), 643–646 (2006).

[20] D. Battaglia and M. Rasetti, *Quantum-like diffusion over discrete sets*, Physics Letters, A. 313, 8–15 (2003).
[21] M. Harada and G. Wilkin, *Morse theory of the moment map for representations of quivers*, Geom. Dedicata. 150, 307–353 (2011). *Preprint* arXiv math.DG 0807.4734v3.)
[22] G. Harder and M. S. Narasimhan, *On the cohomology groups of moduli spaces of vector bundles on curves*, Math. Ann. 212, 215–248 (1974/5).
[23] M. Gromov, *Structures métriques pour les variétés Riemanniennes*. Paris: Conception Edition, Diffusion Information Communication, Nathan, (1981).
[24] K. Fukaya, *Hausdorff convergence of Riemannian manifolds and its applications*, Advanced Studies Pure Mathematics. 18(1), 143–238 (1990).
[25] J. Ambjørn, M. Carfora and A. Marzuoli, *The geometry of dynamical triangulations*, Lecture Notes in Physics. New York: Springer-Verlag, (1997).
[26] G. Wilkin, *Homotopy groups of moduli spaces of stable quiver representations*, Int. J. Math. (2009). *Preprint* arXiv 0901.4156.
[27] P. Diaconis, K. Khare and L. Saloff-Coste, *Gibbs sampling, exponential families and orthogonal polynomials*, Statistical Science. 23(2), 151–178 (2008).
[28] C. N. Yang and R. Mills, *Conservation of isotopic spin and isotopic gauge invariance*, Phys. Rev. 96(1), 191–195 (1954).
[29] T. Tao, *What is a gauge?* (2008). Available at http://terrytao.wordpress.com/2008/09/27/what-is-a-gauge/.
[30] M. F. Atiyah and R. Bott, *The Yang-Mills equations over Riemann surfaces*, Philos. Trans. Roy. Soc. London Ser. A. 308(1505), 523–615 (1983).
[31] C. P. Rourke and B. J. Sanderson, *Block bundles: I, II, III*. Annals Math. 87(1) 1–28, (2) 256–278, (3), 431–483 (1968).
[32] O. Knill, *The Dirac operator of a graph*. (2013). *Preprint* arXiv math.CO 1306.2166v1.
[33] W. V. D. Hodge, *The theory and applications of harmonic integrals*. Cambridge: Cambridge University Press, (1941).
[34] B. Farb and D. Margalit, *A primer on mapping class group*. Princeton: Princeton University Press, (2011).
[35] J. Baeten, *The history of process algebra*, Theoretical Computer Science. 335(2–3), 131–146 (2005).
[36] H. H. Lewis and C. H. Papadimitriou, *Elements of the theory of computation*. New Jersey: Prentice-Hall, (1998).
[37] J. D. McCarthy and A. Papadopoulos, *Simplicial actions of mapping class groups*, in *Handbook of Teichmüller Theory Vol. III* (A. Papadopoulos ed.). Zürich: European Mathematical Society Publishing House, 297–423 (2012).
[38] W. P. Thurston, *Three-dimensional geometry and topology*. Princeton: Princeton University Press, (1997).
[39] M. Rasetti, *Is quantum simulation of turbulence within reach?* International Journal Quantum Information, (2014). *In press* DOI: 10.1142/ S0219749915600084.
[40] A. Asok, B. Doran and F. Kirwan, *Yang-Mills theory and Tamagawa numbers: the fascination of unexpected links in mathematics*, Bull. London Mathematical Society. 40(4), 533–567 (2008).
[41] W. Crawley-Boevey, *Geometry of the moment map for representations of quivers*, Compositio Math. 126(3), 257–293 (2001).
[42] D. Zagier, *Elementary aspects of the Verlinde formula and of the Harder-Narasimhan-Atiyah-Bott formula*, Israel Mathematical Conference Proceedings. 445–462 (1996).

[43] E. M. Gold, *Complexity of automaton identification from given data*, Information and Control. 37, 302–320 (1978).
[44] J. Baeten, F. Corradini and C. A. Grabmayer, *A characterization of regular expression under bisimulation*, Journal of ACM. 54(2), 1–28 (2007).
[45] L. Mosher, *Mapping class groups are automatic*, Ann. of Math. 142(2), 303–384 (1995).
[46] V. Lapshin, *The topology of syntax relations of a formal language*. (2008). *Preprint* arXiv math.CT 0802.4181v1.
[47] M. Artin, A. Grothendieck and J. L. Verdier (eds.), *Théorie des topos et cohomologie étale des schémas Séminaire de Géométrie Algébrique du Bois Marie* 1963-64, (SGA 4) Vol. 1 Lecture notes in mathematics (in French). 269. Berlin: Springer-Verlag, xix, 525 (1972).
[48] E. Merelli, M. Pettini and M. Rasetti, *Topology driven modeling – the IS metaphor*, Natural Computing. (2014). *In press* DOI: 10.1007/s11047-014-9436-7.
[49] C. Barrett, H. B. Hunt, M. V. Marathe, S. S. Ravi, D. J. Rosenkrantz, R. E. Stearns and M. Thakur, *Predecessor existence problems for finite discrete dynamical systems*, Theor. Computer Sci. 386, 3–37 (2007).
[50] H. Haken, *Information and Self-Organization: a macroscopic approach to complex systems*, Series in Synergetics. New York: Springer-Verlag, (2010).
[51] J. A. S. Kelso, *Dynamic patterns: the self-organization of brain and behavior*. Cambridge: The MIT Press (1995).
[52] E. Frenkel, *Lectures on the langlands program and conformal field theory*. (2005) *Preprint* arXiv hep-th math.AG math.QA /0512172v1.
[53] I. Runkel, J. Fjelstad, J. Fuchs and C. Schweigert, *Topological and conformal field theory as Frobenius algebras*, Contemp. Math. 431, 225–248 (2007).
[54] S. Rees, *Hairdressing in groups: a survey of combings and formal languages*, in *Geometry & Topology Monographs* Vol. 1: The Epstein Birthday Schrift (I. Rivin, C. Rourke and C. Series, eds.). International Press, 493–509 (1998).
[55] D. B. A. Epstein, J. W. Cannon, D. F. Holt, S. V. F. Levy, M. S. Paterson and W. P. Thurston, *Word processing in groups*. Boston: Jones and Bartlett (1992).
[56] M. R. Bridson and R. H. Gilman, *Formal language theory and the geometry of 3-manifolds*, Comment. Math. Helv. 71, 525–555 (1996).
[57] G. Petri, P. Expert, F. Turkheimer, R. Carhart-Harris, D. Nutt, P. J. Hellyer and F. Vaccarino, *Homological scaffolds of brain functional networks*, J. Roy. Soc. Interface. 11 20140873 (2014). DOI: 10.1098/rsif.2014.0873.