The Topological Field Theory of Data: a program towards a novel strategy for data mining through data language

2015 J. Phys.: Conf. Ser. 626 012005

(http://iopscience.iop.org/1742-6596/626/1/012005)

Journal of Physics: Conference Series **626** (2015) 012005 doi:10.1088/1742-6596/626/1/012005

The Topological Field Theory of Data: a program towards a novel strategy for data mining through data language

M Rasetti [1] **and E Merelli** [2]

[1]ISI Foundation, Via Alassio 11-C, 10126 Torino (Italy)
[2] School of Science and Technology, University of Camerino, Via del Bastione 1, 62032 Camerino (Italy)

E-mail: `rasetti@isi.it ; emanuela.merelli@unicam.it`

Abstract. This paper aims to challenge the current thinking in IT for the 'Big Data' question, proposing – almost verbatim, with no formulas – a program aiming to construct an innovative methodology to perform data analytics in a way that returns an automaton as a recognizer of the data language: a *Field Theory of Data*. We suggest to build, directly out of probing data space, a theoretical framework enabling us to extract the manifold hidden relations (patterns) that exist among data, as correlations depending on the semantics generated by the mining context. The program, that is grounded in the recent innovative ways of integrating data into a topological setting, proposes the realization of a Topological Field Theory of Data, transferring and generalizing to the space of data notions inspired by physical (topological) field theories and harnesses the theory of formal languages to define the potential semantics necessary to understand the emerging patterns.

1. The landscape

Complex Systems are ubiquitous; complex, multi-level, multi-scale systems are everywhere: in Nature, but also in the Internet, the brain, the climate, the spread of pandemics, in economy and finance; in other words in Society. A deep, intriguing question that has been raised about complex systems is: 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?

The challenge is indeed enormous. In statistical mechanics a number of basic restrictive assumptions play a crucial role: *ergodicity* ensures that all accessible states of a system are reached with equal probability; the *thermodynamic limit*, $N \to \infty$, induces the number of particles into play, measured in terms of the Avogadro's number, to be essentially infinite; *particles* are *identical* and *indistinguishable*, constraint that is not even mentioned when studying the features of collections of particles, but it is there – particles of the same species are identical and interact with each other pairwise all in the same way, that is, with the same interaction law – in the quantum case they are indistinguishable; *analytical structure* can be defined for the *underlying dynamics*, that is, regular equations of motion exist at the micro-scale – analyticity breaking and singularities only appear as signal of the macro-phenomenon of phase transitions; *'experiment-based' phenomenology* is repeatable, as in reductionist science, under the same

initial and boundary conditions the same experiment should give the same outcome.
Only under these conditions statistical mechanics is able to provide a reliable representation of most of the facts of thermodynamics.

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

Today our world is no doubt complex and data-based: about 4 billion people own a cell phone (which makes this the first device in human history owned by more than a half of the world population), 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 zettabyte (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 on the last item of the list first: data; indeed Big Data. The challenge is to extract the huge amount of information, as a norm hidden, flowing in and around complex systems.

Big Data have a variety of diverse features; they are present in science (see, e.g., the Hubble and Genoma projects, the CERN data archive, etc.): typically well organized in high quality data-bases; but now also in society, where they might 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: the Big Data hardware challenge (high performance computing); the Big Data manipulation challenge: both in computer science (new computing paradigm; interaction-based computing; beyond Turing machine) and data analytics (new approach for data mining; non-linear 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 ICT becomes an integral part of the fabric of nature and society.

Goal of the game is to endow ICT with new, more and more efficient tools to play its role in the difficult process of turning *data* into *information*; information into *knowledge*; and eventually knowledge into *wisdom*. In other word, to contribute for a new computing paradigm able to manage the dialectic relations between structural and functional properties in the same way as the human brain interacts with information and behaves as a set of embodied computers.

Aim of this manuscript is to explore the possibility of taming Big Data with Topology (the geometry of 'shapes'), bearing 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 (geometrically represented) in which information is encoded; the frame for algorithmic (digital) thinking; the lode where to perform data mining, that is, to extract patterns of information. It generates the goal as well: finding new ways of mining data spaces for information resorting to geometrical – indeed topological and combinatorial – methods.

First of all we claim – extending ideas proposed by Carlsson, Edelsbrunner and others to deal with the problem of image reconstruction – that geometry and topology are the natural tools to handle large, high-dimensional, complex spaces of data. *Why*? Because: *global, though qualitative information* is relevant: the data user aims to obtain knowledge, i.e., to understand how data is organized on large scale. *Metrics are not theoretically warranted*: in physics, most phenomena naturally support elegant, clear-cut theories which imply exactly what metrics to use; in the life or social sciences this is much less cogent. *Coordinates are not natural*: data are typically conveyed and received in the form of vector-like object (strings of symbols, typically numbers in some field), but the 'components' of these vectors have no meaning as such and their

linear combinations are not objects in data space. In other words, the space of data is not a vector space and those properties of data space that depend on a specific choice of coordinates cannot be considered relevant. *Summaries are most valuable*: the conventional method of handling data is to build 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 some given η; then try and determine the optimal choice of η. The complete diagram that illustrates the arrangement produced by hierarchical clustering is however much more informative, as it captures at once the summary of all relevant features under all possible values of η. The problem here is to try and find a way to know how the global features of data space vary varying η.

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 space, namely, connectivity, classification of loops and higher dimensional manifolds, and invariants, that are, properties which are preserved under homeomorphisms of the ambient space. *Topology* studies geometric properties in a way less sensitive to metrics than geometry, it ignores the value of distance functions and replaces it just with a notion of proximity ($\eta \approx$ connective 'nearness'). *Topology* deals with those properties of geometric objects that do not depend on coordinates but only on intrinsic geometric features; it is *coordinate-free*.

Moreover, in topology, relationships involve maps between objects; thus they are a manifestation of *functoriality* and, what is more subtle and deeper, the invariants are related not just to objects, but to maps between objects as well. Functoriality reflects then a *categorical* structure, allowing for computation of global invariants from local information.

Last, typically full information about topological spaces is faithfully contained in their *simplicial representation*, a piece-wise linear (PL), combinatorially complete, discrete realization of functoriality. As already recalled, 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. Such graph, however, whilst it captures well data connectivity (local), it ignores a wealth of higher order (global) features, well discerned instead thinking of it as the scaffold (1-skeleton) of a higher-dimensional object: the *simplicial complex*. The latter is a PL space built from simple pieces (simplices) identified combinatorially along their faces, obtained by completion of the graph.

2. The challenge

Current thinking in Information Technology is at the crossroad of different evolution pathways. On the one hand, is what is now universally referred to as the 'Big Data' question [1], 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. On a different side is the ever increasing number of practical instances, in science as well as in the analysis of societal issues, of problems that require going computationally 'beyond Turing' [2], which touch on decidability and computability, as well as on embodied computation. These bear on the recently Cerf's raised question of whether or not there is any real 'science' in computer science [3], namely if all the well posed questions can be approached by a truly scientific methodology, universal and self-contained.

Following Cerf's view, we assume modeling of computational processes as the most inspiring candidate for the construction of a true 'theory', able to sustain predictions about complex processes through the analysis of large data sets. We acknowledge also Pearl's diagrams [4] – surprisingly analogous to Feynman's representation of interactions in quantum field theory, with cause-effect relations replacing time flow direction – aiming to construct analytic equations that not only fully characterize a problem of this sort, but make its solution accessible. Main point

at issue is indeed to eliminate unimportant details while revealing the underlying structure: a method well known to statistical mechanics (renormalization group), dealing with fluctuations and noise induced by interactions, and to chaos theory (dynamical effects of nonlinearity), where patterns emerge despite the apparent randomness of the processes.

Present paper intends to describe a program designed to suggest a possible novel way 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 high-level query languages that allow insignificant details to be suppressed and information to emerge as 'mined out' correlations. Principal goal of the paper is the definition of a theoretical framework, described essentially as a non-linear topological field theory, as 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: *information as element of a language*.

The approach proposed differs from all others for its deep roots in the inference of globally rather than locally coded data properties. Its focus is on the integration of constructive elements of topological data analysis (i.e., facts as forms) into a topological field theory for the data space (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 described. A sort of *language of forms* recognized by a proper automaton closely related to the field theory.

Incidentally, but not unexpectedly, this has a profound, far reaching philosophical meaning: in the words of Wittgenstein [5], " The world is the totality of facts, not things. ... The facts in logical space are the world. ... The logical picture of the facts is the thought." and " To imagine a language means to imagine a form of life. ... The meaning of a word may be defined by how the word can be used as an element of a language game."

Outcome of the approach proposed will be a way to discover directly from the space of observations (the collection of data) those relations that encode (by mens of a language) emergent features in the complex systems; *patterns* that data describe as correlations among events at the global level, result of interactions among systemic components at local level. Complex systems global properties are hard to represent and even harder to predict, just because – contrary to the information conventional science deals with – complex systems knowledge is typically based not on repeatable experiments and shared phenomenology that, incidentally, provide information about the statistical features of the system, but on *data* or on virtual artificial representations of real systems built out of data. Emergent features of complex systems arise out of these interactions; whilst the global patterns they create react back on those low lying levels, with no need of developing tools to sustain a learning process.

Tree bodies of knowledge, that are the three pillars our scheme rest on, need to operate synergically: i) *Singular Homology Methods*, tools for the efficient (re-)construction of the (simplicial) topological structures which encode patterns in the space of data; it enables to make *Topological Data Analysis* – homology driven – resting on the global topological, algebraic and combinatorial architectural features of the data space, equipped with an appropriate 'measure'; ii) *Topological Field Theory*, a construct mimicking physical field theories, to extract the necessary characteristic information about such patterns in a way that – in view of the field non-linearity and self-interaction – might generate as well, as feedback, the reorganization of the data set itself; it supports the construction of *Statistical/Topological Field Theory of Data Space*, as generated by the simplicial structure underlying data space, an 'action', a suitable gauge group and a corresponding fibre (block) bundle; iii) *Formal Language Theory*, a way to study the syntactical aspects of languages - the inner structure of patterns - and to reason and understand how they behave; it allows to map the semantics of the transformations implied by the non-linear field dynamics into automated self-organized learning processes. The three pillars interlaced in such a way as to allow us to identify structural patterns in large data sets and

efficiently perform there 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.

3. Topological data analysis

Main pillar of the whole construction is the notion of *data space*, the crucial feature of which is that it is neither a metric space nor a vector space[1], but is a topological space. This is at the very root of all crucial questions at the basis of the scheme proposed: whether the higher dimensional, global structures encoding relevant information can be efficiently inferred from lower dimensional, local representations; whether the necessary reduction process (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 space global structure; whether the process can be carried over in a truly metric-free way [7]; whether such global topological information can be utilized to extract *knowledge* as well as correlated information, in the form of patterns in data space.

The basic principles of the approach stem out of the seminal work of a number of authors: G. Carlsson [8], H. Edelsbrunner and J. Harer [9], A.J. Zomorodian [10], and others. Fundamental goal of it is to overcome the conventional method of simply converting the collection of points in data space into a combinatorial graph $\mathcal{G}$ (a *network*); an object encompassing all relevant local topological features, whose edges are determined by some given notion of $'$proximity$'$, characterized by a parameter η somehow fixing 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 *global* nature. Such features 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*. As such, $\mathcal{S}$ is built from higher and higher dimensional simple pieces (simplices) identified (combinatorially) along their faces. It is this operation that makes the subtlest features of the data set, seen eventually 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 η, the set of relations induced by η itself as a measure of proximity. In other words, each hyperedge is a *relational* simplex, i.e., a simplicial complex built by gluing together lower-dimensional relational simplices that satisfy the η property. This makes η effectively metric independent: in fact an n-relation here is nothing but a subset of n related data points satisfying the property represented by η. 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 η, namely filtering relations by their robustness with respect to η, 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, qualifying them as candidates to be considered as *signal*, while those that have short-lived features can be assumed to characterize *noise*. Moreover, it defines the notion of an η-driven semigroup connecting spaces in the ensemble.

Key ingredients of this analysis are the homology groups, $H_i(X)$, $i = 0, 1, \ldots$, of X and in particular the associated *Betti numbers*, a basic set of topological invariants of X, the i-th Betti number, $b_i = b_i(X)$, being the rank of $H_i(X)$. Intuitively, homology groups are a functional algebraic tool easy to deal with (as they are abelian) to pick up the qualitative features of a topological space represented by a simplicial complex connected with the existence of *i-holes*

[1] This is unfortunately still acritically assumed by even the most distinguished authors (see, e.g., the book by Hopcroft and Kannan [6]).

(holes in i dimensions) in X. Holes simply mean cycles which don't arise as boundaries of higher-dimensional objects. This is why the number of i-dimensional holes, b_i, is realized by the dimension of $H_i(X)$, quotient vector space of the group of i-cycles with the group of i-boundaries. In the torsion-free cases knowing the b_i's is equivalent to knowing the full space homology. Efficient algorithms are known for the computation of homology groups [11]. Persistent homology is generated recursively, beginning with a specific complex characterized by a given $\eta = \eta_0$ and constructing from it the succession of chain complexes and chain maps for an increasing sequence of values of η, say $\eta_0 \leq \eta \leq \eta_0 + \Lambda$, for some Λ. Complexes grow with η, thus such chain maps can be naturally identified with a sequence of successive inclusions.

Whilst most invariants in algebraic topology are difficult to compute efficiently, homology is somewhat exceptional not only because its invariants arise simply as quotients of finite-dimensional spaces but also because some of its features can be possibly derived from $'$physical$'$ models. Indeed, whereas in standard topology invariants were *constructed* out of geometric properties to distinguish between manifest global, homeomorphically invariant objects, it was shown in physics that other invariants could instead be *discovered*, based, e.g., on topological quantum field theory technology [12]. Such invariants provide information about purely topological properties one cannot detect, not even hint, based on purely geometric representation.

It is this latter perspective that is adopted here, namely the idea of constructing a reliable $'$physical$'$ scenario for data spaces, where no structure is visible a priori. $'$Physical$'$ should of course be interpreted metaphorically. We want to figure out a coherent formal framework in an abstract space of data, where no equation is available, to describe through its topology the hidden correlation patterns coding data into information, in a way analogous to that one follows, say, in general relativity, where a given distribution of masses returns the full geometry of space-time. Here we expect that information hidden in data should 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.

This touches of course on how patterns are to be interpreted, as it bears rather on pattern discovery than on pattern recognition. In logic there are approaches 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 exactly that kind of structural regularity, namely organization of configurations or symmetry, that one identifies with the notion of *correlations* in (statistical) physics [13]. These logical paradigms will guide our strategy.

A delicate issue is here that simplicial complex $\mathcal{S}$ (typically but not automatically a finite 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. 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 $\dim \mathcal{S} \geq 5$ there is only the surgery obstruction to $\mathcal{S}$ being homotopy equivalent to a closed manifold, and if $\mathcal{S}$ is homotopy equivalent to a manifold then the complex behaves as the base space of a unique Spivak normal fibration[2]. This entails that all finite simplicial complexes have

[2] A (smooth) manifold has a unique tangent bundle and a unique stable normal bundle; but a finite Poincaré complex does not possess such a unique bundle. Nevertheless, it possesses a substitute, in some sense spherical, fibration that is unique: the Spivak normal fibration. This has a property 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, whose fibre is homotopically equivalent to a sphere.

the homotopy type of manifolds with boundary.

An additional observation is that available algorithms to compute persistent homology groups all focus on the notion of filtered simplicial complex, consisting of pairs built out of the simplex obtained at each given step in the recursive construction and the order-number of the step, the time-like discrete ordering parameter (label) at which that simplex appears in the filtration; a picture that can be naturally interpreted as the representation of a *process* endowed with a proper inherent dynamics, similar to a discrete-time renormalization group flow [15]. One may then expect, in analogy with dynamical triangulations of simplicial gravity, that the combinatorially different ways in which one may realize the sampling of *inequivalent* structures in the persistence construction process, varying the simplex structure, give raise to a $'$natural$'$ probability measure. The measure thus generated is consistent with the data space invariants and transformation properties.

4. From topology to field theory

Besides the customary filtrations due to Vietoris-Rips [16], whose k-simplices are the unordered $(k+1)$-tuples of points pairwise within distance η, 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 in which to implement persistence, another filtration needs to be considered, entering here naturally into play, Morse filtration. In the case of simplicial complexes that are manifolds, this is a filtration by excursion sets, in terms of what for differentiable manifolds would be curvature-like data; here it is a non-smooth, discretized, intrinsic, metric-free version thereof, appropriate for the wild simplicial complex that is data space, that can be obtained by the simplicial, combinatorial analog of the Hodge construction.

Even though apparently dealing with metric-dependent features Morse filtration is indeed purely topological. Morse theory generates a set of inequalities for alternating sums of Betti numbers in terms of a corresponding alternating sum of the numbers of critical points of the Morse function for each given index. Also, simplicial Morse theory generates intrinsic, discrete notions of gradient vector field and gradient flow associated to any given Morse function f_M. The latter, particularly significant if interpreted in the framework of discrete differential calculus, had applications in classical field theory over arbitrary discrete sets [19], described as well in a non-commutative geometry setting [20].

The crucial feature here is that a Morse complex, built out of the critical points of a Morse function with support on the vertices of $\mathcal{S}$, has the same homology as this underlying structure. This assumes particular significance, because the induced Morse stratification, [21], is essentially the same as the Harder-Narasimhan [22] stratification of algebraic geometry, thus one can construct the analog of local $'$co-ordinates$'$ around the Morse strata and represent the negative normal bundle to the critical sets. What relates Morse with homology is the property that the number ν_i of critical points of index i of a given function f_M is equal to the number of i cells in the simplicial complex structure obtained $'$climbing$'$ f_M, that bears on b_i. Noticeably, Morse homology, which can be defined using a generic f_M and an induced, local, topologically invariant metric (i.e., one independent of the function and on the metric itself), is isomorphic to the singular homology: in other words, Morse and Betti numbers encode the same information, yet Morse numbers allow us to think of and underlying $'$manifold$'$.

Inspired by what happens in the context of gravity, Gromov-Hausdorff (GH) [23, 24] topology is selected to construct a self-consistent measure over $\mathcal{S}$. Gromov's spaces of bounded geometries provide in fact the natural framework for addressing the questions posed by high-dimensional simplicial geometry; in particular for establishing entropy estimates that characterize the distribution of combinatorially inequivalent simplicial configurations (a problem solved in gravity theory [25], where however it is much easier, as one deals there with an underlying metric

vector space that gives rise, under triangulation, to a simplicial complex which is a Lorentz manifold). This choice leads naturally to the construction of a statistical field theory of data, as the statistical features of GH topology are fully determined by the *homotopy* types of data space [26]. Both complexity and randomness of the emerging structure can be quite large in the case of big data, since the number of $'$coverings$'$ of a simplicial complex of bounded geometry is expected to grow exponentially with the volume. A $''$thermodynamic limit$''$ needs then to be built over the growing filtrations of simplicial complexes, that look more and more random. As a well defined statistical field theory requires dealing 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, this amounts to including the property that the substrate underlying the Gibbs field may itself be in some way random. One needs therefore to resort 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. Notice that this may lead to critical behavior, as diversified phase structures emerge – entailing a sort of phase transition when the system passes from an homotopy type to another. A scenario emerges: the deep connection between the simplicial complex structure of data space and the information that such space hides, encoded at its deepest levels, resides in the property that data can be partitioned in a variety of equivalence classes classified by their homotopy type, all elements of each of which encode similar information. In other words, in X information behaves as a sort of $'$order parameter$'$.

5. A Field Theory of Data

A single mathematical object may be assumed to encompass most of the information about the global topological structure of the data space: $\mathcal{P}(z)$, the Hilbert-Poincaré series (indeed 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 $'$template$'$ 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 the connection field over a manifold M (a Riemann surface), and the gauge group G, under which the Chern-Simons' (CS) action (i.e., the $(2k\text{-}1)$-form defined in such a way that its exterior derivative equals the trace of the k-th power of the curvature) is invariant, is $SU(N)$.

To begin with, let us recall – paraphrasing Terry Tao [29] – that a *gauge* is simply a global $'$coordinate system$'$ that varies depending on one's location with respect to the reference space. A gauge transformation is a change of coordinates consistently applied to *each* such location, and a gauge theory is the model for a system to which gauge transformations can be applied letting dynamics unchanged. A global coordinate system is nothing but an isomorphism between some geometric or combinatorial objects in a given class and a standard reference object in that same class, as opposed to a local coordinate system, in which the analogous isomorphism is between local pieces of objects. In a coordinate-invariant perspective all geometric quantities must be converted to the values they assume in that specific representation, so that every statement is invariant under coordinate changes. If this can actually be done, the theory can be cast into 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. Every coordinate system arises in this manner. Thus, the space of coordinate systems can be identified with the isomorphism group G of the standard object. This group is the *gauge group* for the class of objects considered. Such general definition will allow us to still think of $'$coordinates$'$; however, not as one does for vector spaces but simply as an intrinsic way to identify mutual relations between objects.

Let us continue the analogy, yet stressing out the differences. The arena for YMFT is a generic, smooth manifold, M, over which the connection field is well defined and allows for a consistent definition of the action (the curvature is simply the exterior derivative of the connection plus the wedge product of the connection by itself). Field equations are then nothing but a *variational* $'$machine$'$ which takes as input a symmetry constraint, expressed as invariance with respect to G, and produces 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 on M a vector space, i.e., a *fiber* over that point: the field at p is simply an element of such fiber. The resulting collection of objects (manifold M plus fibers at every point $p \in M$) is a *vector bundle*. In the presence of a gauge symmetry, the fiber must be the copy of a representation of the corresponding group G. In other words, 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, and Harder and Narasimhan [22], via a purely combinatorial approach, have established a powerful formula expressing the Hilbert-Poincaré series (global) in terms of those corresponding to all Levi subgroups of G as a functor of the YMFT, in 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 much more complex, because of the more complex underlying structure. Let us however recall that vector bundles, proper to the differential category, have a PL category analogue, referred to as *block bundles* [31]. These allow us to reduce geometric and transformation problems over manifolds to homotopy theory for the various groups and complexes involved. This provides a natural way to reconstruct the moduli space of G-bundles in a discretized setting. The construction needs to be extended to a simplicial complex that, as already noticed, in general is not a manifold. Fortunately, Novikov's lesson is that this can be done in homotopy terms [14]. Since the homotopy class of a map fully determines its homology class, the simplicial block-bundle construction clearly 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 gives just the Poincaré series [32]. $\mathcal{K}$ can be obtained by constructing over the simplicial complex an intrinsic (metric-free) combinatorial Laplacian [33]. This is done by the *ad hoc* construction of the Hodge decomposition over $\mathcal{S}$ and the related Dirac operator.

Regarding 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, but are constrained by the manipulation processes they can be submitted to – there is only one natural symmetry it needs to satisfy: invariance under all those transformations of data that don't change its topology and are consistent with the constraints. This implies that the gauge group be the semidirect product $\mathcal{G} \wedge \mathfrak{G}_{\mathfrak{MC}}$ of the group $\mathcal{G}$ associated with the characteristic process algebra of the data set 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 showing behavior, which is the total of events or actions that it can perform, together with the order in which they are executed and other aspects of this execution, such as timing or probabilities, that define the process. The term algebra refers instead simply to the fact that the approach taken to represent behavior is algebraic and axiomatic. That is, operations on processes are defined, and their effects are represented formally in terms of methods and techniques of universal algebra.

In analogy with the definition of a group as any mathematical structure consisting of a single universe of elements, with operators on this universe of elements that satisfy the group axioms, a *process algebra* is any mathematical structure satisfying the axioms given for its operators,

and a process is then an element of the universe of this process algebra. The axioms allow calculations with processes. Process algebra has thus its roots in universal algebra, however it often goes beyond the strict bounds of universal algebra: the restriction to a single universe of elements can be relaxed, e.g., different types of elements can be used, and sometimes binding operators are considered. It supports mathematical reasoning about behavioral equivalences that, independent of the specific approach followed for its definition, must be a congruence with respect to behavioral operators.

In automata theory, on the other hand, a process is modeled as an automaton. An automaton has a number of *states* and of *transitions*, namely, ways of going from a state to its 'neighbor' states through the execution of elementary actions, the basic units of behavior. Besides, an automaton has an initial state (possibly, more than one) and a number of final states. A behavior is an execution path of a number of elementary actions that leads from the initial state to a final state. Given this basic behavioral abstraction, an important issue is to decide when two automata can be considered equal, expressed by a notion of equivalence, the *semantic equivalence*. In automata, the basic semantic equivalence is language equivalence: an automaton is characterized by its set of execution paths, and two automata are equal when they have the same set of execution paths. In this context, an algebra that allows reasoning about automata is the algebra of regular expressions [36].

In the automata model the notion of 'interaction' is missing, yet during the execution from initial state to final state, a system may interact with another system. When dealing with models representing interacting systems, 'concurrency theory' is used: the theory of interacting, parallel, distributed or reactive systems. Process algebra is an algebraic approach to concurrency theory, and as such it may have parallel composition among its basic operators. In this case, automata are 'transition systems' for which the notion of equivalence is not necessarily restricted to language equivalence. Prominent among the equivalences is *bisimilarity*, which considers two transition systems equal if and only if they can mimic each other's behavior in any state they may reach.

Finally, 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) 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). Moreover, 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 $\mathcal{A}$.

As for $\mathfrak{G}_{\mathfrak{MC}}$, the actions of (extended) mapping class groups on spaces of different sorts encoding geometric and topological objects, like homotopy classes, foliations, conformal structures have been extensively studied [37]. These actions are all induced from the actions of homeomorphisms of the base space on the corresponding objects. Moreover, the spaces on which the mapping class groups act can be equipped with various 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 present context are actions by simplicial automorphisms on the different 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 we fully recover through the construction proposed the whole data space topology (for example the set of Betti numbers) the space of data, but we are able to continue 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.

6. The Formal Language Theory facet

The construction outlined so far naturally brings to light a new facet, as much unexpected as elegant in its form: a Formal Language Theory (FLT) dimension.

A preliminary question to raise is then whether the topological landscape adopted is inherently coherent with the structure of 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, 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 $\mathcal{G} \wedge \mathfrak{G}_{\mathfrak{MC}}$. In other words, such languages are each a collection of discrete spaces containing a finite number of homeomorphic objects. In other words, by the TDFT we construct indeed a language of data, the language proper to topological shape $\mathcal{S}$.

Interpreting the gauge group G as topological shape language requires to 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 able to distinguish whether or not a given word, representation of a group element, is in ′canonical form′, and hence if two elements in canonical form differ and by which generators. It may be worth recalling that automatic groups were actually originally introduced in connection with topology, in particular with the study of the fundamental group, and hence of the homotopy, of 3-manifolds, because the class of automatic groups can be extended to include the fundamental group of every compact 3-manifold 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 formal language theory 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 of the notion of grammar, defined by the relations 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 resorting to the notion of syntax diagram, $\mathbb{D}$, that is the connected multigraph with nodes labeled in terms of the formal languages 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 first, for each $\mathbb{D}$, the collection of subdiagrams labelled 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, labelled by some symbol of $\mathfrak{A}$, it includes some neighborhood of this symbol. Such 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] (the Grothendieck topology is a structure on a category that makes its objects act like the open sets of a topological space). It is the formal language generated by the field theory through its gauge group that makes the TDFT consistent with a formal language architecture.

7. Language, structure and behavior, automata

On a different front, the notions the TDFT construct outlined have crucial consequences in terms of information theory. In particular, the three basic identifications: the *architectural structure* as a *fiber bundle*, consisting of a base space, here the space of data X, viewed as a topological space, and a fibre attached to each point of it; the *fibers*, as each a representation of the *gauge*

group $G = \mathcal{G} \wedge \mathfrak{G}_{\mathfrak{MC}}$; the *field* as an element of the *fiber* at each point of the data space; an *action* – here, in the simplest non-interacting case, the combinatorial Laplacian – able to describe the variations of the global topological landscape, have a far reaching interpretation.

Such architecture is indeed what allows us to touch the final goal: the definition of a methodology whereby, starting from the exploration of (large) sets of data one could extract a $'$language$'$ capable of describing them as a unified system of structure and behavior. This new object can be interpreted as a *true* (effective, extended) *data space* including, besides the topological features inherent in the data set, the set transformations allowed among them, generated by the group of all its possible topology-preserving transformations and reflected in the ensuing equivalence classes. In such perspective, the system is itself a *self-organizing* $'$program$'$, whose identifiers are the interactions that the field action implies. These interactions correlate parts of potential processes (embedded programs) of real life applications: a characteristic feature caught in the $S[B]$ paradigm [48].

The principle of *self-organization* has long entered as fundamental feature the theory of nonlinear (possibly discrete [49]) dynamical systems. It provides the clue to represent the diverse pictures of the relation between lower level elements and higher order structures in a complex system. 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 [50, 51] to induce their emergence. 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.

In an algebraic prospective the language signature becomes a measure of the interactions that generate the environment associated with the data set. This is exactly what is done 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*). That is equivalent in TFTD to determine which is the fiber bundle associated to a given element of a fiber attached to a given point of the topological space.

The structure of $S[B]$ can be naturally identified as a fiber bundle: $S[B] = \big(B, S, \pi, \mathcal{B} \doteq \{B_j \,|\, j \in \mathcal{J}\}\big)$, with total space B, base space S, projection map $\pi : B \to S$ and fiber set $\mathcal{B}$. $\mathcal{J}$ is a label set tagging points $s_j \in S$, $j \in \mathcal{J}$. In $\mathcal{B}$ each single fiber B_j specifies the global topological constraints conditioning all the correlations of s_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, for any subset of constraints corresponding to a given choice of the global invariants, the inner homeomorphisms among equivalence classes on the fiber.

An important, though simpler, analog of this construction comes from conformal field theory (CFT), that it somehow mimics. 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 category $\mathfrak{T}$ (the so called Modular Tensor Category). 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 $\mathfrak{M}$ 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, articulated construction a crucial notion emerges: that of (asynchronously) $\mathfrak{L}$-*combable* group [54], namely a group to each element of which can be associated a word in

some free group in 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 transformations (flows) of the data set. The class of combable regular languages consists of precisely those groups that are asynchronously automatic. Recalling 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 totally contained in the structure of the monoid), which is typically finite. In this context regular languages are referred to as languages of $'$data words$'$. A rigorous, but simple construction of data words consists in identifying first alphabet $\mathfrak{A}$, 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 such languages, whose basic feature is that they provide a trade-off among three important properties: *expressivity*, good *closure* properties and decidable (or efficiently decidable) *emptiness*; striking an acceptable balance in this 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 iff 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.

The relevant emerging relationships naturally involve, in the topological setting, invariants are related not only to objects but maps between pairs of objects as well. This is an explicit manifestation of *functoriality*. Functoriality is central in algebraic topology as it permits computation of global homological invariants from local information, a feature that reflects the inherent, natural categorical structure of this tool. The theory of automata and formal languages merge with the field theoretical picture just in this way, because the field theory generates sequences of symbols, that enter into play in the simplicial construction of the G-bundle associated with with the gauge group G, and relations among them. This bears on the enumerative combinatorics content of the theory, because G is reduced essentially to homotopy braids, that provide the language recognized by the automata. Moreover, 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 set is finite and bounded, independently of $\mathfrak{L}$.

The use of formal languages leads as well 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 very large data space, and to manage, query and reason over those concepts so as to infer new knowledge. This is remindful of 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. Yet, while 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 yet 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 the hidden structural aspects, 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 once 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.

8. Patterns

The conclusive step, that results by merging all the above ingredients into a unique, consistent field-theoretical picture, in a 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, bears on the individuation 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 though delicate. Use of the characteristic patterns of the data system 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 such a perspective, 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, saying little about what that pattern is: and this is 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, 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 strategic directions as how to search data space. Whilst several details of the theory remain to be exhaustively worked out, its grand design – just presented – at least programmatically does not, except perhaps in some of its subtlest technicalities, 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 'psychedelic state' based on functional magnetic resonance imaging data [7, 57].

9. Conclusions

A few final remarks. We have outlined the construction of a topological gauge field theory able to act 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 the output is pattern sets in the form of field correlations as generated by the field equations. Such 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 simplicial 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.

Acknowledgments
The financial support is acknowledged of the Future and Emerging Technologies (FET) programme 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] Various Authors 2011 Special Section: Dealing with Data *Science* **331**
[2] Barry Cooper S 2012 Incomputability after Alan Turing *Notices AMS* **59** (6) 776–784
[3] Cerf V 2012 Where is the science in computer science? *Commun. ACM* **55** (10) 5
[4] Pearl J 2009 *Causality: Models, Reasoning, and Inference* (Cambridge: Cambridge University Press)
[5] Wittgenstein L 1921 Logisch-Philosophische Abhandlung *Annalen der Natur und Kulturphilosophie, Leipzig* **14** 184–262
[6] Hopcroft J and Kanna R 2013 *Foundations of Data Science* (http://blogs.siam.org/the-future-of-computer-science/)
[7] Petri G, Scolamiero M, Donato I and Vaccarino F 2013 Topological Strata of Weighted Complex Networks *PLoS ONE* **8** (6): e66506. DOI: 10.1371/journal.pone.0066506
[8] Carlsson G 2009 Topology and data *Bull. AMS* **46** (2) 255–308
[9] Edelsbrunner H and Harer J 2010 *Computational Topology, an Introduction* (Providence: Amer. Math. Soc.)
[10] Zomorodian AJ 2009 *Topology of Computing* (Cambridge: Cambridge University Press)
[11] Basu S, Pollack R and Roy M-F 2006 *Algorithms in Real Algebraic Geometry* (New York: Springer-Verlag)
[12] Witten E 1989 Quantum field theory and the Jones polynomial *Commun. Math. Phys.* **121** (3) 351–399
[13] Shalizi CR and Crutchfield JP 2001 Computational Mechanics: Pattern and Prediction, Structure and Simplicity *J. Stat. Phys.* (3-4) **104** 816–879
[14] Novikov SP 1996 On manifolds with free abelian fundamental group and applications *Izv. Akad. Nauk SSSR* ser. mat. **30** (1) 208–246 – English translation: 1967 *A.M.S. Transl.* **67** (2) 1–42
[15] Zinn-Justin J 2002 *Quantum field theory and critical phenomena* (Oxford: Clarendon Press)
[16] Vietoris L 1927 Über den höeren Zusammenhang kompakter Räume und eine Klasse von zusammenhangstreuen Abbildungen *Math. Ann.* **97** 454–472
[17] Čech E 1932 Théorie générale de l'homologie dans un espace quelconque *Fund. Math.* **19** 149–183
[18] de Silva V 2003 A weak definition of Delaunay triangulation (*Preprint* arXiv cs.CG/0310031 v1)
[19] Auchmann B abd Kurz S 2006 A geometrically defined discrete Hodge operator on simplicial cells *IEEE Trans.* MAG-**42** (4), 643–646
[20] Battaglia D and Rasetti M 2003 Quantum-like Diffusion over Discrete Sets *Phys. Lett.* A **313** 8–15
[21] Harada M and Wilkin G 2011 Morse Theory of the Moment Map for Representations of Quivers *Geom. Dedicata* **150** 307-353 (*Preprint* arXiv math.DG 0807.4734v3)
[22] Harder G and Narasimhan MS 1974/75 On the cohomology groups of moduli spaces of vector bundles on curves *Math. Ann* **212** 215–248
[23] Gromov M 1981 *Structures métriques pour les variétés Riemanniennes* (Paris: Conception Edition, Diffusion Information Communication, Nathan)
[24] Fukaya K 1990 Hausdorff convergence of riemannian manifolds and its applications 1990 *Advanced Studies Pure Math.* **18** (1)
[25] Ambjørn J, Carfora M and Marzuoli A 1997 *The Geometry of Dynamical Triangulations* (Lecture Notes in Physics Springer-Verlag)
[26] Wilkin G 2009 Homotopy groups of moduli spaces of stable quiver representations, *Int. J. Math.* in press (*Preprint* arXiv 0901.4156)
[27] Diaconis P, Khare K and Saloff-Coste L 2008 Gibbs Sampling, Exponential Families and Orthogonal Polynomials *Statistical Science* **23** (2) 151–178
[28] Yang CN and Mills R 1954 Conservation of Isotopic Spin and Isotopic Gauge Invariance *Phys. Rev.* **96** (1) 191–195
[29] Tao T 2008 (http://terrytao.wordpress.com/2008/09/27/what-is-a-gauge/)
[30] Atiyah MF and Bott R 1983 The Yang-Mills equations over Riemann surfaces *Philos. Trans. Roy. Soc. London* Ser. A **308** (1505) 523–615

[31] Rourke CP and Sanderson BJ 1968 Block Bundles: I, II, III *Annals Math.* **87** (1) 1–28, (2) 256–278, (3) 431–483
[32] Knill O 2013 The Dirac operator of a graph (*Preprint* arXiv math.CO 1306.2166v1)
[33] Hodge WVD 1941 *The Theory and Applications of Harmonic Integrals* (Cambridge: Cambridge University Press)
[34] Farb B and Margalit D 2011 *A primer on Mapping Class Group* (Princeton: Princeton University Press)
[35] Baeten J 2005 *The history of process algebra* **335**(2-3), 131-146
[36] Lewis HH and Papadimitriou CH 1998 *Elements of the theory of computation* (Prentice-Hall)
[37] McCarthy JD and Papadopoulos A 2012 Simplicial actions of mapping class groups, in Papadopoulos A (Ed.) *Handbook of Teichmüller theory* Vol. III (European Mathematical Society Publishing House, Zürich) 297–423
[38] Thurston WP 1997 *Three-Dimensional Geometry and Topology* (Princeton: Princeton University Press)
[39] Rasetti M 2014 Is quantum simulation of turbulence within reach? *Int. J. Quant. Inf.* (In press DOI: 10.1142/S0219749915600084)
[40] Asok A, Doran B and Kirwan F 2008 Yang-Mills theory and Tamagawa numbers: The fascination of unexpected links in mathematics *Bull. London Math. Soc.* **40** (4), 533–567
[41] Crawley-Boevey W 2001 Geometry of the moment map for representations of quivers *Compositio Math.* **126** (3), 257–293
[42] Zagier D 1996 Elementary aspects of the Verlinde Formula and of the Harder-Narasimhan-Atiyah-Bott Formula *Israel Mathematical Conference Proceedings* 445–462
[43] Gold E. M. 1978 Complexity of Automaton Identification from Given Data *Information and Control* **37** 302-320
[44] Baeten J, Corradini F and Grabmayer CA 2007 A characterization of regular expression under bisimulation *Jou. of ACM* **54** (2)
[45] Mosher L 1995 Mapping Class Groups are Automatic *Ann. Math.* **142** (2) 303–384
[46] Lapshin V The topology of syntax relations of a formal language *Preprint* arXiv math.CT 0802.4181v1
[47] Artin M, Grothendieck A and Verdier J-L (Eds.) 1972 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
[48] Merelli E, Pettini M and Rasetti M 2014 Topology driven modeling – the IS metaphor *Natural Computing* (In press DOI: 10.1007/s11047-014-9436-7)
[49] Barrett C, Hunt III HB, Marathe MV, Ravi SS, Rosenkrantz DJ, Stearns RE and Thakur M 2007 Predecessor existence problems for finite discrete dynamical systems *Theor. Computer Sci.* **386** 3–37
[50] Haken H 2010 *Information and Self-Organization: A Macroscopic Approach to Complex Systems* (New York: Springer-Verlag, Series in Synergetics)
[51] Kelso JAS 1995 *Dynamic patterns – The self-organization of brain and behavior* (Cambridge: The MIT Press)
[52] Frenkel E 2005 Lectures on the Langlands Program and Conformal Field Theory (*Preprint* arXiv hep-th math.AG math.QA /0512172v1)
[53] Runkel I, Fjelstad J, Fuchs J and Schweigert C 2007 Topological and conformal field theory as Frobenius algebras *Contemp. Math.* **431** 225–248
[54] Rees S 1998 Hairdressing in groups: a survey of combings and formal languages in: *Geometry & Topology Monographs* Vol. 1: The Epstein birthday schrift 493–509
[55] Epstein DBA, Cannon JW, Holt DF, Levy SVF, Paterson MS and Thurston WP 1992 *Word processing in groups* (Boston: Jones and Bartlett)
[56] Bridson MR and Gilman RH 1996 Formal Language Theory and the Geometry of 3-Manifolds *Comment. Math. Helv.* **71** 525–555
[57] Petri G, Expert P, Turkheimer F, Carhart-Harris R, Nutt D, Hellyer PJ and Vaccarino F. 2014 Homological scaffolds of brain functional networks *J. Roy. Soc. Interface* **11** 20140873 DOI: 10.1098/rsif.2014.0873