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