systems

MDPI

Article

Big Data: From Forecasting to Mesoscopic Understanding. Meta-Profiling as Complex Systems

Gianfranco Minati

Italian Systems Society, 20161 Milan, Italy; gianfranco.minati@airs.it; Tel.: +39-02-6620-2417

Received: 1 September 2018; Accepted: 28 January 2019; Published: 8 February 2019

Abstract: We consider Big Data as a phenomenon with acquired properties, similar to collective behaviours, that establishes virtual collective beings. We consider the occurrence of ongoing non-equivalent multiple properties in the conceptual framework of structural dynamics given by sequences of structures and not only by different values assumed by the same structure. We consider the difference between modelling and profiling in a constructivist way, as De Finetti intended probability to exist, depending on the configuration taken into consideration. The past has little or no influence, while events and their configurations are not memorised. Any configuration of events is new, and the probabilistic values to be considered are reset. As for collective behaviours, we introduce methodological and conceptual proposals using mesoscopic variables and their property profiles and meta-profile Big Data and non-computable profiles which were inspired by the use of natural computing to deal with cyber-ecosystems. The focus is on ongoing profiles, in which the arising properties trace trajectories, rather than assuming that we can foresee them based on the past.

Keywords: collective behaviour; dynamics; fuzzy; mesoscopic; meta-profile; model; non-computable

1. Introduction

This article's purpose is to contribute to the introduction of fresh understandings of and approaches to Big Data to allow multiple comprehensions that are suitable for the realisation of clues and tendencies that are appropriate for conjecture, identify processes, properties, and forecasts. As we introduce later, Big Data are not only considered to be representative of well-defined phenomena [1], but also as having autonomously acquired properties that are detectable at the mesoscopic (or topological) level, such as collective behaviours [2].

We consider the possibility of different definitions compared to the usual ones under which Big Data are collected and without the use of an a priori hypothesis [3], like the classic definition of belonging for sets. However, Big Data are collections of such size and dynamical variety that classical analytic approaches are unsuitable.

New understandings can aid in the development of more complex research and modelling, in addition to searching for properties considered hidden by the large amount of data and undergoing detection with, for example, statistical and data mining methods [4]. Basically, it is a matter of sophisticating the research strategy with new hypotheses. For example, we move away from the vision of IBM data scientists, where Big Data is broken into four dimensions: volume, variety, velocity, and veracity [5].

We continue to consider Big Data as large amounts of data that are not based on any model, but we introduce some possible conceptual variations, for example, Big Data as a digital accumulation of untreated waste from other processes that are not currently considered due to the usual 'no data deletion' concept. Furthermore, we consider Big Data not as representative of a phenomenon, but as a phenomenon to be possibly studied, which has autonomous properties similar to collective behaviours.

We assume that dealing with Big Data as a phenomenon contains not only traces of past events, but also the assumption of having properties that are independent from the original event, when studied as a virtual collective being [6]. Thus, we consider the occurrence of ongoing, non-equivalent, multiple properties with any initial temporal beginning, duration, and variable combination in conceptual correspondence with the nature of quasi-systems, which occur when a system is not always the same system and or even a system at all [7,8].

We then mention the criticality of the level of description, as well as the scalarity assumed to detect the properties as well as the conceptual ineffectiveness of increasing the quantity of data available. Before proposing new approaches allowed by the conceptual 'definitions' mentioned above, we consider the differences between models and profiles. Models (especially ideal) are expected to support understanding, while profiling is intended to represent non-ideal, data-driven models, and their emergent, ongoing properties, for example, correlations and coherences [9]. Another differentiation considered is between forecasting and understanding. Forecasting is based on the importance of the past, is based on analogies and repetitiveness, and takes some contextual conditions into account. Understanding is considered to be the ability to conjecture, hypothesise, speculate, and, finally, realise the nature of the phenomenon under study, for instance, chaotic, hosting bifurcations, fluctuations, and the presence of multiple dynamic coherences.

We specify that the differences introduced above should be considered in a constructivist conceptual framework in which, instead of trying to find how a phenomenon 'really' is, one looks for the most effective way to think of it (multiple modelling) [10]. For instance, we consider the conceptual framework of the theoretical physical principles of uncertainty, complementarity, indeterminacy, and epistemological incompleteness [11] of the so-called Dynamic Usage of Models (DYSAM), which is suitable when the complexity of the phenomenon under study is such that different and non-equivalent models should be simultaneously applied, like for the usage of logical openness [12,13].

This also relates to the Subjective Theory of Probability introduced by De Finetti. Probability is not intended to be objectivistic and observer-independent. Probability is computed based on the expectancies and configurations of events considered by the observer. The past has little or no influence, since any configuration of events is new and the probabilistic values to be considered are reset [14].

Furthermore, we consider the differences between *dynamics* and *structural dynamics*. In the first case, dynamics refers to well-studied dynamical systems modelled by time-dependent variables and parameters considered by dynamical systems theory, such as analytical models, the concepts of equilibrium, limit cycles, and chaos.

We consider *structural dynamics* to be changes within the systems under study that do not relate *only* to variables and are parametric of the *same* temporal analytical representations, that is, the *same* structure ([8], pp. 87–117). Structural dynamics considers sequences of both structural and non-structural changes together with the properties of such sequences, as in collective behaviours. Examples include the occurrence of coherence among properties of *sequences* of singularities, such as phase transitions and singularities.

Based on what was discussed and introduced previously, we introduce methodological and conceptual proposals, using both mesoscopic variables and their properties to profile and meta-profile (infra properties) Big Data, as well as non-computable profiles inspired by natural computing dealing, for instance, with ecosystems and machine learning.

We finally mention a possible *mesoscopic understanding of Big Data* that is suitable for the realisation of *tendencies* and *ongoing properties*.

The focus is on ongoing profiles, when they emerge, and the properties of their emergence that trace the trajectories rather than assuming that we can foresee the past within Big Data. It is a matter of *an ongoing mesoscopic understanding* that is equivalent to *mesoscopic profiling*.

We mention the approaches considered by the Topologic Data Analysis (TDA) in this meta-analysis-based framework [15].

We conclude by mentioning some possible related future research.

2. Big Data

IBM researchers estimated that around 2.5 quintillion (10^{30}) bytes of data are generated daily. Such data are generated, for instance, from cell phone signals; digitalised pictures and videos; financial transactions, purchase transaction records (in *real time economy* as well where purchases and sales are made by algorithms, for example, in stock exchange); posts to social media sites; email messages and attachments; sensors used to gather environmental information; in security systems and traffic controllers; and in words typed in documents, printed in newspapers and books or transmitted through telephone networks and broadcasts [1]. These data are named *Big Data* [16–18]. Within this data, we must distinguish between those that are publicly available and those that are unavailable for any reason, such as confidentiality or technical reasons. We should consider the *quality* of Big Data, for example, their reliability, non-duplication, completeness, and spatial and temporal alignment.

Furthermore, we should also consider how reliable the spatial and temporal alignments of data are to allow for consistent comparability and classification. On the other hand, the quantity of the data available and its redundancy may allow for the predominance of some properties, even in fuzzy ways, that is sufficient to make consistent inferences.

Having said this, we move on from both considering Big Data only as large amounts of data without any a priori assumptions and from the concept of the four Vs [5]. In this section, we consider Big Data in different ways. For instance, it is considered to be digital, accumulated, untreated waste generated by processes that are not currently considered and by the usual 'no data deletion' which has the effect of indefinite accumulation. Furthermore, we consider Big Data as an autonomous phenomenon that acquires rather than possesses properties for, for example, emergent collective behaviours and phase transitions. We mention the crucial role of the level of scalarity and description at the mesoscopic level. We consider multiple properties and their infra-properties among mesoscopic properties like correlations and coherence.

2.1. Big Data as Large Amounts of Data Not Based on Any Model

On the other hand, we may consider the different, but in some ways similar, case of the availability of microscopic data related to a phenomenon where the data are unable to be explained or modelled. The fact is that it is impossible [18,19] to model, reconstruct, nor deduce significant semantic [20] levels from such data. Contrary to what was assumed by the reductionist illusion, *it is eventually only possible to make constructivist ex post hypothesises*. *More is not better*.

Examples include microscopic, molecular information of *material* makeup, for instance, machines, electronic devices, and biological living entities. The increase in the amount of data is useless since the data are meaningless, *per se,* if not coupled with hypotheses or theories. In this regard, we limit ourselves to only mentioning how some authors [21] have introduced the topic relating to the possibility of theory-less knowledge being replaced by suitable concordances, correspondences, and correlations within Big Data. However, "...the problem relates to a cognitive research strategy to produce knowledge. There are more and more cases of knowledge produced without the search for, or availability of, theories, by using concordances and correspondences in a data deluge often termed Big Data [18], using data-driven approaches within very large databases" [22].

However, the assumption that "correlation supersedes causation and theorising" has been demonstrated to be *mathematically wrong* [18].

Other cases include *ecosystems* (see note 7; in our case, *cyber ecosystems*), for which microscopic information is not available, while it is possible to obtain macroscopic information, for example, the temperature and volumetric measurements for ponds. A similar case is given by the collective behaviours (such as flocks) for which macroscopic information is obtainable, for example, the average speed, altitude, and direction, while only in particular cases is the microscopic information available (such as when the constituting agents are equipped with Global Positioning System (GPS), for example, car traffic and herds). Other cases include astrophysical data detected from the universe for which microscopic information is almost undetectable and intractable, for example, from biological and chemical systems of molecules, networks of neurons, or atmospheric or geological phenomena (for example, lava eruptions) [23].

2.2. Big Data as Accumulated on Treated Digital Waste

We refer here to the enormous collections of continuously produced waste. We may consider items of any kind, such as compost, glass bottles, packaging, paper, plastics, and sewage treatment plants.

There are various stages of collection.

In the absence of treatment and recycling, there is an accumulation of garbage, such as depositions in possibly temporary landfills. In the latter case, some microscopic information is maintained or recoverable, for instance, traces of the packed and bottled material; information written on paper, for example, the language, handwriting, and dates; and plastic types. Some direct reuses are also possible, such as for deposits for car wreckers and junk dealers.

We also consider the ideal cases of shredder treatments for documents and cryptography, which produce *virtual waste* by reducing the reversibility as much as possible, although it is theoretically still possible.

After the very initial stage of waste collection, while maintaining the microscopic nature of the items, the following accumulation stages and treatments irreversibly increase the macroscopic nature, leading to the loss of origin information such as temporal information, information related to usage, geographical information, and correlational information. The focus is rather on the material properties of the waste to be processed.

However, the second macroscopic aspect also adds some acquired cumulative information such as the daily amount of selected waste produced and the geographical temporally accumulated information.

Metaphorically, Big Data can be considered *digital accumulated untreated waste*. Such accumulations also allow for some microscopic reversibility, for instance, when dealing with timed documents, such as receipts, tickets, and invoices. *This relates to the usual 'no data deletion' accompanied by an increasing irretrievability over time*.

Treatments considered for Big Data are usually finalised to extract, find correlative information, and verify assumptions. Is it possible to conceptually consider reuses? Possible reuses relate to simulations of scenarios with partial previous Big Data available. Reuses may consider the reoccurrence in similar frameworks of complete scenarios when some partial aspects take place.

2.3. Conceptual Understanding

A very important point is that Big Data does not come after any process, inquiry, or measurement. We do not look for them as specifying models, measurements, parameters, or hypotheses. Big Data are, in some way, self-generated, and they *constitute* the phenomenon. We may model or measure them, *but we need to take into account that the more information they have and the older they are, the weaker and fuzzier the connections with the generating phenomena become. The increase in the quantity, mixture, and time make Big Data acquire autonomous properties such as those of their sequences, for example, correlational, ergodic, periodical, and statistical.*

In our case, the data are the items in, for example, a train ticket, such as the emission time, the price, the place number, and the arrival and departure stations, no matter whether they are recorded on a mobile telephone or printed. We understand Big Data to be a collective entity, a collective being of data (see Section 1), established by dynamical combinations, sequences, superimpositions, and multiple systems of data as *virtual entities*.

Big Data occur with or without possible different temporary materialisations. They *fluctuate* in the sense that they can have different materialisations, limited durations, accumulations, and sedimentations, either because they are not removed or because they are registered as well as purposely reconstructed.

Big Data are intended in the following to describe *virtual phenomena* whose macroscopic properties, such as those related to correlation, equilibrium, and statistics, partially or completely lose correspondence with the microgenerating phenomena.

2.4. Multiple Properties of Big Data as a Virtual Phenomenon

We consider here the alternative approaches to profiling by using measurements and constitutive data, as mentioned above, from the perspective of using the past to foresee the future (for instance, to predict and explain by analogy with the past), and search for lucky cases due to, for instance, the discovery of diffused mathematical regularities and statistics [24,25].

Rather, we consider here the approaches to suitably profile Big Data, understood as large amounts of data constitutive of the phenomenon [26] (that is, Big Data *is* the phenomenon; metaphorically, Big Data is an emergent[1] swarm). The inhomogeneity of the variables may be intended to relate to the same phenomenon as a kind of emergent *collective being* [26], of which we consider different aspects, variables, and measurements.

Rather than relying on assumptions to deal with different temporally changing values presumed to refer to the *same* property, we consider here the occurrence of ongoing, non-equivalent, multiple properties that have any initial temporal beginning, duration, and variable combination.

This is conceptually equivalent to approaches considering the collective behaviours as being established by the occurrence of multiple interactions, each with a different duration and beginning. Multiple interactions may apply to any sequence and involve the same entities belonging, at the same or any time, to different systems (for the concept of multiple systems, see note 3, in [8,26].

We consider the use of the approaches introduced to model collective behaviours to also profile Big Data in regard to the conceptual correspondence between interactions (in collective behaviours) and properties (in Big Data). Accordingly, we propose properties of the mesoscopic variables used to model collective behaviours (real phenomena) to profile Big Data (virtual phenomena of data).

We consider the ongoing aspects as being constructivist choices made by the observer.

2.5. The Role of the Level of Description

In some disciplines, such as physics, it is possible to adjust the level of description when assuming, for instance, a microscopic or macroscopic level. In these cases, general, *formal* adjustment of scaling occurs through the introduction of new variables to allow microscopic information to be ignored and new macroscopic information to be considered.

It is a bit like adjusting the focus of a lens.

A completely different problem is dynamical *adjusting* that looks for meaningful semantic representation(s). For instance, when dealing with *Gestalt figures*, the observer must adjust multiple levels that are being driven by the search for meaning. Another case occurs when dealing with the incompleteness [11] of *Basso Continuo* and *Impressionism*, where the observer *gives cognitive shape* through meaning.

The prospect of determining the structures, meanings, and values of semantic variables seems to be not *easily* (even theoretically?) reducible to procedures and computations [30].

[1] Let us consider the case of populations of *interacting* entities as it is well-known that the process of *interaction* takes place when one's properties and behaviours influence the another's. Interactions may occur though the exchange of energy and information to be elaborated. Inside such interacting populations, different processes may occur. We mention the occurrence of self-organisation, where sequences of properties acquired in a phase transition-like manner have regularities and repetitiveness. As examples, we may consider the repetitiveness of the formation and behaviours in queues in traffic; swarms of mosquitoes around a light; pelicans around stacks of trash; and self or remote synchronisations. Another process is the one intended to establish **emergence** when regularities and repetitiveness occurring for self-organisation are substituted by *coherence*. In this case, sequences dynamically acquire multiple different synchronisations and correlations. We may consider the formation and better emergence of flocks, shoals of fish, and swarms with multiple and changing shapes, densities, and directions as examples. However, over time, they continuously acquire and maintain *coherence, scale invariance, and long-range correlations* in such a way that they become recognisable as the *same* flock, shoal, or swarm [27–29].

The microscopic pulverisation is such that the microscopic data are almost all semantically equivalent.

However, when dealing with a large amount of data that have some *homogeneity*, it is possible to use statistical approaches and data mining procedures.

Techniques of data mining are also available as processes of *discovering* patterns in large datasets. Topological Data Analysis (TDA) is an approach to the analysis of large datasets using techniques from topology [31].

Metaphorically speaking, disordered datasets given by words of a story or notes of a symphony seem intractable from this point of view when looking to resume their coherences to represent meaning.

When dealing with Big Data, we face the *conceptual ineffectiveness* of increasing the quantity of data available, leading to one of two possibilities, that is, the rise of redundancy or the rise of noise; neither of these converge to give meaning.

3. Models and Profiles

We first consider some introductory differences between models and profiles that occur even if there are areas of conceptual overlap, for example, when using random and fuzzy variables.

In the first case, we consider ***models***, in short, as variables and their structures, for example, equations suitable to not just *imitate*, but represent and *understand* the phenomenon under study (distinction introduced by Turing, see, for instance [32,33]. Understanding allows simulation, but this is not the case for the converse. On the other side, the *incompleteness* [11,30] of a model relates to the non-explicit representation (for example, the impossibility of *zipping everything* into a model of equations and using *ideal models* given by general principles assumed to be valid for any phenomenon); to the need to use non-equivalent models simultaneously; to the validity of uncertainty principles (accuracy in measuring one variable is at the expense of another); the validity of principles of complementarity (for example, wave particles); the establishment of singularities (for example, defects); and the need to operate in the context of *logical openness* characterised by the *theoretical* unavailability (because of multiple non-equivalences) of a complete and explicit model of the system and its interactions with what is considered its environment ([8], p. 16, [34], pp. 47–51).

Examples of model properties occur when following the same analytical representations. *It is a matter of possessing the same analytical properties*.

In contrast, we may consider *non-ideal models* (conceptually close to profiles) to be intended as mixtures of general principles and specific choices, for instance, data-driven approaches that are statistically and retrospectively clustered.

We have experienced very large databases in different disciplinary fields. Usually, there are stable, well-defined configurations of sources assumed to be generators of data when the data relate to observables, variables, and measures that are specified by the *observer* and the *observation process*. Research is able to introduce *new* sources corresponding to new models, as used to happen historically; for instance, in physics, when changing models from mechanical to thermodynamic, electromagnetic, optical, and quantum.

In previous cases, researchers considered data as being related to specific phenomena and their models.

Models come first. Data populate models after measurements.

In the second case, we consider ***profiles***—in short, as data of different *natures* that are not related to models.

In the case of profiles, we consider data-driven, *emergent, ongoing properties, for example, correlations and coherences*—both properties of data *representing* real events and properties of data *constituting* the phenomenon, such as the case of Big Data considered above.

It is a matter of searching properties within the available data. The data are possibly ordered *collections* of *classifiable* information. Classifications take place, for instance, because of presumed *homogeneity*, for example, measures of the same variable, and because of *properties*, for example, the 'same' correlation and ranges of validity. Correlations are data-driven, that is, not explicitly stated through fitting with models and analytical representations.

Related possible approaches are given by *data mining* [4], which is devoted to the *discovery* of embedded, supposedly *hidden* patterns through, for instance, suitable computerised data analyses, for example, cluster analysis. The assumption is that the presence of *formal regularities* is hidden by their possible irregular combinations, randomness, and unsuitable scaling.

Examples of traditional profiles in databases include patterns, correlations, and the statistical properties of the data under study [35–39].

Moreover, other approaches include using the network properties of networked available data and using fuzzy logic, as mentioned below. For profiling, *it is a matter of possessing the same non-analytical properties*.

Data come first. Properties should be recognised. The main reason for profiling phenomena is to allow for structural forecasting.

4. Constructivist Approach

The situation is reminiscent, in some ways, of some constructivist understandings about research. It is well-known that *objectivism* relates to research as an attempt to find the truth, the reality as it is, and to *answer* questions and *solve* problems, that is, to find *the* solution ([8,12], pp. 193–194).

This seems to be, at least, an elementary strategy that is unsuitable for dealing with complexity where processes of emergence occur that require the use of multiple non-equivalent models to allow the establishment of new, unexpected, coherent properties (see the case of DYSAM below).

Within the *constructivist* conceptual framework, we stress how experiments are understandable, like *questions* that are metaphorically posed to nature. In turn, metaphorically, nature *responds* by making the experiments happen. However, continuing the metaphor, we get no answers if there are no questions.

Furthermore, events may turn into answers if we invent the appropriate questions to which the answers may be considered. Objectivism is just a particular case ([6], p. 6) where it is assumed that the questioning occurs in steps and *converges* at the truth and not as a multidimensional evolutionary process of non-equivalent knowledge.

When dealing with Big Data, we may consider two generically corresponding approaches: trying to find and *discover* the hidden properties possessed by Big Data or trying to invent questions and question Big Data.

In the first case, we implicitly assume that hidden regularities will be discovered, such as the fractal nature, Fibonacci rules, and regular distributions. Statistical approaches, data mining, and Topological Data Analysis are examples of related searching approaches.

In the second case of constructivist nature, *questioning Big Data*, we take an *abductive* approach. The concept *of abduction* was introduced by Charles Sanders Peirce (1839–1914) as a process of forming explanatory hypotheses. Furthermore, abduction is considered to be a logical operation that is able to introduce any new idea, i.e., non-equivalent to previous ones [40]. We also mention Foerster [41], who considers that *anomalies* (e.g., strange, unexpected behaviours and irregularities) in the environment are not objective, but rather are given by the inappropriateness of the approaches, concepts, and models used to understand the phenomenon under consideration. Abduction is intended to be the cognitive process of hypothesising, inventing, and formulating new models that are suitable for "normalising" what were previously considered anomalies. Furthermore, abduction can be considered as the adaptation, selection, and multiple usage of the most suitable options that are already available.

We consider the constructivist approach in the conceptual framework, for instance, of the theoretical physical principles of uncertainty, complementarity, and indeterminacy, and the epistemological incompleteness of the Dynamic Usage of Models (DYSAM). This is suitable for cases where the complexity of the phenomenon under consideration is so high that the use of different non-equivalent models is necessary[2] (see ([6], pp. 64–75) and there is need for *logical openness* [12,34,42].

In conceptual correspondence with the invention of experiments as questions, both to inquire and to interpret the available data, as considered by constructivism, we propose a modus operandi that combines ideal and non-ideal approaches.

We mention how this approach is consistent with considering the observer as the generator not of simple relativism, but rather of the *cognitive reality*, such as the hypotheses, models, and configurations that we consider to be probabilities, as introduced by De Finetti (1906–1985) [14,43]. This is taken into account in the following section.

5. Foreseeing and Understanding in Structural Dynamics

Following the previous sections, our understanding of Big Data and the role of the level of description was presented, the differences between models and profiles were mentioned, and the principles of the constructivist approach considered in the following text were detailed; we now take into account the differences, if any, between forecasting and understanding.

The possible differences are considered here when differentiating between dynamics and structural dynamics.

In the case of dynamics, we refer to the well-known properties of dynamical systems and modelling by time-dependent variables and parameters that are considered by dynamical systems theory, such as analytical models, the concepts of equilibrium, limit cycles, and chaos [44,45]. Such systems have been studied by dynamical systems theory ([9,26], pp. 64–65).

Rather, our interest is on *structural dynamics*, where changes occurring within the systems under study do not relate *only* to the variables and parametric characteristics of the *same* temporal analytical representations (simplified: to the *same* structure). Structural dynamics considers sequences of both structural and non-structural changes together with the properties of such sequences. Structural dynamics is considered with reference to complex systems ([6], pp. 64–87).

As an example, we may consider *sequences* of phase transitions occurring on the same *materiality*. Generic structural dynamics is intended to be given by such sequences. An interesting case takes place when considering properties of such sequences, for instance, when acquiring coherences for particular complex systems [9].

The establishment of *collective beings* [6] and quasi-systems [8] moving between the statuses of system and non-system and keeping significant levels of coherence is a generic example. Actually, collective beings as collective behaviours emerge from coherent structural changes. Specific examples include anthills, car and signal traffic, flocks, herds, industrial districts, social systems such as cities and markets shoals of fish, swarms, telephone and transportation networks, and termite mounds.

However, the concept of structural dynamics applies to several cases and processes, such as

1. Change, acquisition, loss, and non-linear combinations of structures;
2. Multiple systems[3];
3. Adaptation and learning;
4. Radical changes due to processes of emergence;

[2] An example occurs when considering that a *medical problem* may simultaneously have biochemical and psychological components. Furthermore, economic, political, and sociological aspects of any *social system* are simultaneously present. Another case is given by *business problems* of different nature, such as financial, managerial, and organisational. However, the non-equivalence of models does not imply a lack of interdependence among related effects. We mention the difference with interdisciplinarity where the same approaches and models are applied in different contexts by changing the meanings of the variables (e.g., chaotic behaviour from the climate to economics and the Lotka–Volterra equations from prey–predator systems to generic competition in economy and electronics).

[3] The concept of a multiple system is considered in ([6], pp. 110–137) as set of systems whose components simultaneously belong to more systems. In this case, the set of systems is coherent, i.e., correlated. In the same way, there are multiple networks where the same nodes belong to different and simultaneous networks [45].

5. Changes in properties.

We initially introduced this comment to the discussion about forecasting and understanding since Big Data should tend, by its nature, to have structural dynamics generated by a dynamic variety of phenomena, as mentioned in Section 1.

The ability to *forecast* probably does not necessarily demand the ability to understand. As a matter of fact, the ability to forecast may come from reasoning based on analogies and repetitiveness, and by taking into account some contextual conditions. In the latter case, forecasting is, however, based on the past and not on structural characteristics and properties. Furthermore, the ability to deal with *probabilities* considered in a constructivist way, as introduced by Bruno De Finetti, was mentioned above as a specific case.

Forecasting may be partial, as it refers only to some aspects of the phenomenon and has a probabilistic nature [46].

We should also ask what we are interested in forecasting. In the context of structural dynamics, the forecasts to look for should be structural, such as establishing, increasing, maintaining, and weakening general properties acquired by sequences of structural changes.

Examples include coherence; properties of patterns, for example, topological properties; their similarities; their compatibilities and incompatibilities; predispositions; and dominions. Such properties are very similar to those used to model structural dynamics of collective behaviours, as introduced in Sections 6.3 and 6.4.

As for forecasting, we consider the special cases of structural dynamics and the subjective theory of probability. Accordingly, we consider understanding as not being coincident with classical modelling which looks for the best, most effective one. Rather, we consider multiple and non-equivalent approaches when using approaches such as DYSAM and logical openness (as introduced above), and the theoretical incompleteness for non-ideal models considered above are used when it is not possible to *zip everything* into a single model of equations.

Following the above, we can say that we consider the *structural prediction* capability.

We consider, for instance, the *understanding* of the possible *chaotic nature* of phenomenon, which the high dependency on the initial conditions. The possible complex nature of the phenomenon helps us consider the occurrence of *bifurcations* (changes in the topological structure of the system and the number or type of attractors), *symmetry breaking* (when the form of evolution equations remains invariant after a transformation, e.g., rotation, but the form of their solutions changes), and the role of *fluctuations* (deviations of the actual time evolution from its average evolution within a system subject to random forces). In particular, fluctuations may induce catastrophic consequences within systems that have critical points, as determined by the Self-Organised Criticality (SOC) display of scale invariance, where examples include geological phenomena.

Such an understanding helps to realise and hypothesise compatibilities and incompatibilities and equivalences and non-equivalences, and identifies evolutionary modalities that are able, in this case, to eliminate impossible evolutions and circumscribe possible developments.

In sum, we may tentatively say that the ability to understand aids in the ability to structurally forecast, and the ability to forecast aids in (in a minor way) the ability to structurally understand.

We concentrate on profiling Big Data for such structural understanding and forecasting in Section 6.

6. Profiling Structural Dynamics

In this section, we present approaches and comments which converge to a final methodological proposal that has different levels of possible implementations.

In Section 6.1, we outline the differences between top-down and deductive profiling.

In Section 6.2, for the benefit of the following discussion, we introduce the concept of mesoscopic levels of representation and clustering and discuss the related concepts and approaches that are used to foresee and understand collective behaviour.

In Section 6.3, we mention approaches based on *mesoscopic variables* to model the structural dynamics of collective behaviours.

In Section 6.4, we consider the properties of mesoscopic variables as profiles. We then propose considering the properties of *multiple mesoscopic variables* as properties of populations of profiles (with the meta-profile as the profile of profiles).

Finally, in Section 6.5, we propose the approaches for non-computable profiles, for example, learning through Artificial Neural Networks (ANN), which is used for ecosystems (ecosystems of profiles in our case). This will further be elaborated in future research as explained in Section 7.

6.1. Top-Down and Deductive Profiling

In traditional approaches, a first very important methodological distinction is between the following:

- *Top-down profiling* (supervised learning) from available data. This involves testing, for instance, hypothesised correlations and statistical properties. It is a method of quantifying how many entities respect the property. This is related to ideal modelling.
- *Deductive profiling* (unsupervised learning) to detect, by suitable techniques (for example, data mining), not yet hypothesised properties when exploring a database. It is a method of generating hypotheses, for instance, through correlations and the general peculiar properties of specific entities. This is related to non-ideal modelling. *As we mentioned, the usual approaches have the purpose of finding the profiles as data properties that are able to represent categories*. This is related to the *deduction* of profiles.

Although it is possible to consider profiles as a dynamic mixture, by using a combination of the two previous approaches (both top-down, for example, selecting the variables, properties, and thresholds to be considered), and deducing the properties from the data available, for example, through data mining, statistics, and correlations, we focus on bottom-up approaches, as even phenomenological aspects related to the behaviour of complex systems clearly have a bottom-up nature.

6.2. Mesoscopic Levels of Representations and Clustering

As introduced below, we consider the possibility of identifying suitable mesoscopic variables (introduced in Section 6.3) and their properties for the constructivist findings within the data available regarding *emergent ongoing properties, for example, coherences.* These are considered as profiles.

To introduce the concept of the profiles that we have in mind, we must first remind ourselves of some necessary concepts: microscopic, macroscopic, mesoscopic, and fuzzy.

The definitions are given below:

- *Microscopic*: Focusing on single, indistinguishable, equivalent entities, for example, atoms, generic passengers, customers, and members of flocks or swarms;
- *Macroscopic*: Ignoring microscopic properties and focusing on global properties, for example, dimension, shape, density, and temperature. This is assumed by any aggregation of microscopic entities, for example, the volume, weight, and temperature of a glass of water and dimensions, such as the weight and temperature of a billiard ball, regardless of its molecular composition;
- The *mesoscopic* level: Unlike the macroscopic level, this level does not completely ignore the microscopic level, but rather considers only *some* of the microscopic properties available that are suitable to be quantitatively *clustered*. Examples include cars in traffic that cannot increase their speed (we consider cars standing still in the queue, cars slowing down, and cars running at constant speed in the queue); people standing, uphill, or downhill on the stairs; and the belonging of boids in the same spatial volume (whatever their direction, speed, and altitude are). **We *underline* how the mesoscopic level can be considered a conceptual implementation of fuzziness by replacing degrees of *belonging* with values of aggregation.**

The three levels of descriptions allow for the cor*respondin*g microscopic, macroscopic, and mesoscopic *understanding* in the sense that the reasoning occurs by using the variables of such natures.

The subgroups considered by the mesoscopic levels are *clusters*. We have clustering when aggregation is allowed by suitable criteria and approaches, such as the similarities among the measurements. For instance, the statistical techniques of multivariate analysis [47] are finalised to select the homogeneous elements with respect to the measurements of the same variable, for instance, by optimising the differences in values. Clusters may be eventually *represented* by average values, centroids, and the arithmetic mean of all the values. An example is given by the clusters of elements with similar (that is, a minimum difference) speeds among the different elements.

This clustering is possible through the use of suitable and available computational techniques, such as *k-means* [48], where the objective is to minimise the total intra-cluster variance.

For reasons of comparability, the number of clusters should be obviously constant with regard to time. The number of clusters to be considered may be decided with appropriate computational approaches, such as the *Elbow* and *Silhouette* criteria (see, for instance, [49,50]).

Here, we also consider *fuzzified clustering*, a form of clustering in which each item can belong to more than one cluster [51].

Examples of cluster properties are the number of elements, the distribution of elements within the cluster (for example, close to the min, max, average, or randomly spread out), and the thresholds computed ex post (after procedures of clustering). For each cluster, we consider (for instance, the calculation ex post of percentiles) the min and max values that allow for cluster density and centroids to be computed.

We may consider sequences of the same clustering (for example, per price, per speed, or per age) along the time dimension and consider their properties, such as their distributions, correlations, and statistics.

We may consider the different mesoscopic variables that are constituted of the same elements (for example, customers), though related to different properties and clustered, for instance, by the per amount of expense, the quantity of purchase, or the amount of time taken to purchase (see matrix $M(t)$ in Section 6.4).

6.3. Approaches to the Model Structural Dynamics of Collective Behaviours

At the microscopic level, that is, when considering the properties of the *composing interacting agents*, such as boids, customers, and internet users, the emergent coherence of collective behaviours as complex systems has been modelled in the literature by considering their scale-free correlations, for instance, long-range correlations ([8], pp. 80–87), that occur when the number of correlated elements is equal to the total number of elements. Examples of other properties that have been considered are network properties, power laws, and statistics.

A usual approach may consider the vectors of variables such as $V(t) = [v_1(t)_, v_2(t), \ldots, v_n(t)]$ which, in social systems examples of v_n, include variables such as the number of purchase transaction records using cards, the number of classified telephone calls, the number of internet accesses, the quantity of energy consumption, and the number of tickets for travel services, *all per instant*. In this case, v_n are quantities or numbers suitable for the searching of properties like above.

At the mesoscopic level of description, the meta-structure project ([8], pp. 111–128) is aimed at finding the properties of mesoscopic variables in collective behaviours.

Mesoscopic variables along time are intended as vectors per instant whose scalars contain, for instance, the number of elements or thresholds related to the corresponding clusters.

More precisely, we consider the mesoscopic variable $W_p(t) = [w_{p1}(t)_, w_{p2}(t), \ldots, w_{pn}(t)]$, where p is the property considered. For instance, $p = 1$ and $W_1(t) = [w_{11}(t)_, w_{12}(t), \ldots, w_{1n}(t)]$ relates to the clustering of different numbers of purchase transactions, n identifies the clusters in which the transactions are aggregated by the amount of similarity, and the value of $w_{1n}(t)$ is equal to the number of elements contained in the cluster n or the value of the corresponding thresholds computed ex post.

In addition, the *fuzziness*, as in fuzzy sets, fuzzy logic, and fuzzy systems, is also considered [52–55]. As is well-known in classical set theory, the membership degree of elements can be only *0* or *1*, i.e., an element can belong or not belong to a set. In fuzzy sets, the membership degree of elements varies within the interval *0,1*. A specific fuzzy set is characterised by the membership function. The related fuzzy set theory has found application in several disciplinary fields where it deals with problems with incomplete or imprecise information. Examples include engineering and information theory.

Here, we will consider the mesoscopic *fuzzified clustering*[4] of data where the properties are related to a specific phenomenon to suitably represent the *incompleteness* and possible coherences of the complex phenomenon under study.

This approach is used, for instance, for fuzzy searches and matching, by using search engines on the web. This type of search will be able to find matches even when users only enter partial information or misspell words to be used for the search.

On the other hand, the complexity of the collective behaviours is modelled by using cross-properties of simultaneous multiple mesoscopic variables, related to different aspects such as speeds, altitude, and direction. It is a matter of considering the cross-properties among non-homogeneous clusters (see Section 6.4).

Examples of properties related to cross-correlations and statistics among values are assumed by different mesoscopic variables. Other examples include the occurrence of chaotic regularities, possibly with strange attractors in the reoccurrence of the same cluster properties; correspondences among the properties of the thresholds computed ex post; and possible levels of ergodicity among the configurations of clusters ([8], pp. 111–128).

The ongoing aspect of properties as variants, in addition to those listed in the examples above, is the establishment of different dynamic dominions of validities, such as *local* and *remote* synchronisations and correlations that are probably analytically intractable [8,56].

Collective behaviours are intended as a *source* of *large amounts of data;* however, they must be *coherent*. In this case, the phenomena of emergence are intended to have mesoscopic coherence [2] among the interacting agents when, *"...mesoscopic variables are intended to transversally intercept and represent values adopted by aggregates of microscopic variables. Values of mesoscopic variables are considered to represent the effective application of interaction rules"* ([2], p. 55). In the meta-structure project, we consider the properties of the clustered variables considered to be suitable to model the collective behaviours [2,57].

The purpose of the meta-structure project is to find the properties of collective behaviours that are suitable to

- *Recognise* behaviours, for instance, at different temporal and spatial scales where acquired properties, such as patterns, may not be easily recognisable, and
- *Induce*, if not *prescribe* to, for instance, the Brownian-like (random) behaviour of multiple interacting agents to put on such properties in order to assume coherence(s) (see, for instance, ([8], pp. 122–123). This is of interest, for instance, to facilitate, give start, support, and keep coherence of the collective behaviours of cells in biology, prices in the economy, stock exchange in finance, agents in traffic, and cyber swarms of drones, and herds of robots for both terrestrial, marine, public security, and defence interventions. Examples of actions that are suitable to induce collective behaviour include the inclusions of adequate perturbations, such as environmental perturbations.

On one side, the project has the purpose of modelling and understanding collective behaviours to search for ways to simulate and make available some modifying approaches, as mentioned

[4] A form of clustering in which each item can belong to more than one cluster.

above. On the other side, the project has the purpose of *structurally forecasting* the suitability of modifying approaches.

In the following section, we propose the usage of similar approaches for Big Data.

6.4. Profiles as Properties of Mesoscopic Variables

We may consider the profiling of single aspects such as the interest in books, for example, by author, theme, and typology, like narrative or science; travel, for example, by price, location, and modality, like train or plane; and food, for example, by gastronomic fairs, local cuisine, and events, such as wine and beer events.

When dealing with large amounts of microscopic data generated by collective behaviours (Case 2 in Table 1) any approach presumes coherence(s) to be the phenomenological property of the phenomenon under study. Any computable clustering, for example, speeds, directions, altitudes, metrical distances, topological distances, and related properties, is finalised to represent coherence(s) and their dynamics. It is a matter of finding such properties that *represent* the collective behaviours and *applying* the approaches to induce such behaviour.

Table 1. The profiles and meta-profiles.

Profiling Considered	
Profiles	Combinations of top-down and deductive profiling. Properties of $V(t) = [v_1(t), v_2(t), \dots, v_n(t)]$.
Profiles as properties of *mesoscopic variables*	We may then consider profiles as properties of mesoscopic variables represented by vectors $W_p(t) = [w_{p1}(t), w_{p2}(t), \dots, w_{pn}(t)]$. Related and mixed (since combining top-down and deductive approaches where the researcher is supposed to also invent variables and measurements to be considered) approaches lead to the identification of mesoscopic fuzzified clusters and their infra-cluster features.
Meta-profiling (profiles of profiles) in populations of profiles	We consider populations of inhomogeneous vectors $W_p(t) = [w_{p1}(t), w_{p2}(t), \dots, w_{pn}(t)]$ where p is related to different properties and the number n may be standardised ex post by inserting empty clusters. In this way, we may deal with a resulting matrix $M(t)$ given by the properties of each cluster. Meta-profiles are thus represented by properties of sequences along the time of matrix $M(t)$.
Non-computable profiles: emergent profiles	We consider emergent profiles. For instance, a particular case is given by dynamic non-computable-profiles as any machine learning parameters, weights, and levels that are suitable to make an ANN learn the behaviour of Big Data over time.

We may then consider the profiles as properties of mesoscopic variables represented by vectors $W_p(t) = [w_{p1}(t), w_{p2}(t), \dots, w_{pn}(t)]$. The related and mixed (since combining top-down and deductive approaches where the researcher is supposed to also invent variables and measurements to be considered) approaches lead to the identification of mesoscopic and possible fuzzified clusters[5] and their infra-clusters features, as considered below.

We now consider the properties of *multiple mesoscopic variables* as the properties of populations of profiles (with a meta-profile being the profile of profiles, see below).

At this time, we may consider the populations of inhomogeneous vectors $W_p(t) = [w_{p1}(t), w_{p2}(t), \dots, w_{pn}(t)]$ where p is related to different properties and the number n may be standardised ex post by inserting empty clusters. In this way, we can deal with a resulting general matrix $M(t)$ given by p-properties *per* n-cluster, that is, the number of aggregations per similarity ([8], pp. 120–122).

[5] In this regard, we mention the technique of fuzzy clustering when each entity can belong to more than one cluster [51].

In social systems, examples of *W$_p$(t)* include the following:

$$M(t) = \begin{vmatrix} w_{11}(t)_{,} & w_{12}(t), & \ldots, & w_{1n}(t) \\ w_{21}(t)_{,} & w_{22}(t), & \ldots, & w_{2n}(t) \\ & & \ldots & \\ w_{p1}(t)_{,} & w_{p2}(t), & \ldots, & w_{pn}(t) \end{vmatrix}$$

- Purchasing transaction records using clustered cards, for example, by amount, number of items, or recurring groups of items;
- Purchase and sale transactions for clustered shares, for example, by amount, number of shares, and time of the transaction;
- Telephone calls clustered, for example, by temporal duration, time, call area, and destination area;
- Information exchange (for instance, text documents, videos, and signals) clustered, for example, by size, transaction time, and encoding;
- Internet accesses clustered, for example, by software used, location, time, and duration of access;
- Energy consumption events such as electrical events, clustered, for example, by corresponding outside temperature, working hours, and natural light trends;
- Tickets for travel services, such as surface transport, for example, trains, buses, and subways clustered, for example, by geographical area, time period, and service cost range.

These are all measured per instant.

In addition, when dealing with different clusters represented by *W$_{pn}$(t)*, the researcher may consider the level of belonging by specifying the fuzziness.

The mesoscopic aspect relates to the variables by following constructivist decisions[6] made by the researcher together with the properties and threshold levels of the representation of data to be taken in the count. Afterwards, the mesoscopic fuzzified clustering allows for the detection of cluster and infra-cluster analytical properties. However, the availability of this kind of data is restricted to those that are accessible. For example, for purchase transactions, we have the available amounts, items, and time, but not the age or sex of the client. Properties may be fuzzy *retrieved* within Big Data, such as for fuzzy information retrieval systems (see, for instance, [58]). The searching of profiles should be viewed as questions to Big Data—questions that are, however, formulated with the kinds of data available and are only partially chosen by the researcher. This limits the abductive aspect of the demand, focusing instead on the *choices* of variables and properties to be considered. *The level of quantitative aggregation does not replace; however, the semantic meaning to be abductively realised by the researcher*, that is, the meaning of the profile. **The ongoing aspect is given by the fact that we do not search for different values acquired over time by the *same* profile, but rather, for different multiple profiles that identify different aggregations of the available data. This is represented by the properties of sequences along the time period of the matrix *M(t)*.**

As for probability, keeping De Finetti in mind, objective probability does not exist. In the same way, objective profiles do not exist if they are not in standardised boundary conditions.

It is then possible to consider different possible populations of simultaneous and subsequent matrix profiles *M(t)* of interest that correlate, transform, and re-emerge over time. We may, for instance, speak of systems of profiles (eventually categorised) when the occurrence of one implies an interaction with another because, for instance, they have common or interconnected elements or relationships.

In conceptual generic correspondence within the concept of *meta-data*, that is, data that provides information about other data that are used for several applications, such as for research in the semantic

[6] We specify that the choice of a variable is a property that can be considered as a semantic act by the researcher that is not replaceable by algorithmic approaches, such as statistical aggregation processes that can only be supportive. On the other side, a semantic act is also to give ex post meaning to detected clusters and correlations.

web (see, for instance, [20]), we use the term *meta-profile* for *profiles of profiles*. We use this term to relate to the properties of multiple simultaneous or subsequent profiles. This is intended to be a multiple matrix $M_s(t)$, where s indicates the matrix combination of $W_p(t)$ considered by the researcher.

Populations of profiles, as considered above, may be considered to be systems or quasi-systems [8] depending on whether they display systemic properties, such as the maintenance of possible stabilities, equilibriums, periodicities, patterns, correlations, and remote synchronisation [56]. In this case, we consider the populations of profiles as sets of matrixes $M_s(t)$, where s identifies both the semantic and computational selections.

Such meta-profiling is expected to allow for the detection of the possibility or attitude to establish communities that are considered to have compatible, equivalent profiles; the introduction of communities to subsequent profiles; the results or degeneration of previous profiles; and the merging with other profiles and their negations. Another case is given by the dynamics and related properties, such as fuzziness, of multiple profiles that are valid for each instant. *Lucky cases take place when few dominant profiles are sufficient to represent the meta-profile*.

Meta-profiles are intended to allow for model dynamics, evolutionary compatibilities, and equivalences, and to identify areas (incubators) where specific phenomena may emerge.

We may say that profiling by using mesoscopic variables generically coincides with mesoscopic understanding.

6.5. Non-Computable Profiles

In the previous sections, we considered the case of profiles as properties of incomplete, fuzzy, and non-analytically tractable mesoscopic variables. The particular aspects consisted of assuming computable properties of multiple clusters.

Here, we remind the reader of the well-known *Church–Turing thesis* [59]. This thesis claims that each function that is effectively computable by an algorithm can be calculated by a suitable Turing Machine. In other words, *each algorithm is Turing-computable*. Moreover, two very important related properties of any algorithm are its *completeness* and *explicitness*, that is, the computational process (the program) occurs by sequences of finite numbers of steps and specifies the process's causal chain. However, the processes of emergence are intended to be non-algorithmic, non-procedural processes with regard to their input and output [18]. It is interesting to observe how the behaviours of some particular types of neural network are non-Turing-like. This situation suggests the suitability of constituting a theory of *natural computation* [60].

As considered below, different approaches, such as learning, are necessary, including those based on machine learning.

Phenomenological emergence may be intended to describe the *uniqueness* of coherent subsequent phase transitions [8,30].

The phenomenon of the emergence of constituting ecosystems[7] may be intended to constitute the so-called *natural computation* as a physical process. We now consider how, in the effective phenomenology of emergence (that is, how emergence emerges (its 'mechanism'), the boundary conditions continuously change and require the assumption of a level of observer-dependent description. Specifically, they need to consider the user's cognitive processes, for instance, their abilities to memorise recognise, make logical inferences, and find properties.

[7] An ecosystem should be viewed as a dynamical community of multiple systems and quasi-multiple systems ([8], pp. 166–170) formed by the interaction of a community of organisms with their environment. The richness of the interactions is peculiar as they are based, for instance, on (or given by) adaptation, cognitive (at different levels of complexity) interactions, compensations for missing resources, competition, growth, learning, mutual symbiosis, and reproduction. All of this occurs within a framework of spatial and temporal dynamics, different durations, densities, and uses of the environment. Some models represent aspects such as the well-known Lotka–Volterra, or predator–prey models whose equations describe the dynamics of biological systems where the number of individuals in the population changes over time according to the equations.

In the case of natural computation, its *step-by-step* computation is intended to correspond to step-by-step *frames* of a process of radical emergence. In this case, the *sequence* is *the computation itself*, and it is non-Turing. Rather, it has phenomenological nature and it is possibly repeatable thanks to the *learning* processes of a cognitive complex system. We stress how *repeatability by learning* is conceptually different and theoretically non-equivalent to Turing-like computational iterations [18].

The case of *machine learning* performed by Artificial Neural Networks (ANNs) through supervised or unsupervised learning is an example. In the case where the observer establishes correspondences among specific inputs and outputs, the program (in this case, the ANN) *represent*s (can we say *computes*?) the (machine) *learning* which, from a specific input, will generate the corresponding output.

It is well-known that the representation of the process is non-analytical, and it is performed by changing adaptive networks by specifying weighted links and levels among variables.

This is then an example of a process that can be intended to be non-analytical and non-explicit. Such processes are often termed *sub-symbolic*, whereas they emerge from explicit and complete computational processes (algorithms), such as ANN programs, which are explicit algorithms.

We consider the possible dynamic non-computable-profiles to be the classes of Big Data listed in Section 1, for example, any machine learning parameters, weights, and levels that are suitable to make an ANN to learn the behaviour of Big Data over time [61,62].

More generally, here, we consider the case of *emergent profiles*. Rather than only using populations of interrelated profiles, we may consider the ecosystems of interacting profiles where continuous processes of emergence occur.

6.6. Topological Data Analysis

Within the conceptual framework of the meta-scale analysis introduced above and with a focus at the mesoscopic level, we consider the Topological Data Analysis (TDA) that was already mentioned in Section 2.5.

Tt is well known that topology is the branch of mathematics whose area of study is related to the study of features of shapes and the connectivity of spaces.

TDA is a homology-based theory.

In mathematics, homology was introduced as a methodological approach to the categorisation of holes, that is, to identify classes in a manifold.

In addition to the searching for meta-structures allowing for meta-profiles, as mentioned above, some topological approaches were introduced to identify and model the mesoscopic aspects of complex systems intended to be complex networks.

In particular, such approaches in computational topology are based on *persistent homology* (individualisation of generators of persistent n-dimensional holes. This approach is suitable for the identification of non-local structures, like weighted holes inside a link-weight network texture. In this approach, the properties distinguish weighted networks in two main classes. One is given by small and hierarchically nested holes. The other one exhibits significant and longer persisting inhomogeneities. This approach allows the processes of shape recognition and data discovery to occur within large datasets [63].

Another case relates to *topology driven models* in the framework of the Information Sytems (IS) metaphor when disciplinary research is conceived as part of data science. This is the case when the adaptivity of complex systems is considered to be driven by data. The purpose is to obtain global topological information from spaces of data. In particular, it is of interest to consider the persistent homology and Betti numbers of the phenomenological data [64].

Moreover, another TDA approach we would like to mention is related to the usage of the *persistent entropy* that characterises the environment. The interest in this is because the value of such entropy is highly related to the topological structures of the data.

In this case, the entropy measure is essentially computed by using the persistent Betti barcodes. It is a method that considers the topological properties of complex systems [65].

7. What to Do with Big Data—From Forecasting to Mesoscopic Understanding.

Effective classifications [66] allow for suitable searches. This is the case for web search engines. The order of magnitude is now trillions (10^{18}) of searches per year inside the World Wide Web [67]. However, the available information is not very precise due to the number of search engines available and the marketing importance of the data.

A large variety of proprietary approaches are used.

The problem is in having all the data available without knowing it or without knowing what to do with it.

Network representations allow for the representation of properties and equivalences of paths that can be used as profiles to be possibly considered, dynamically combined and weighted, and used to establish meta-profiles.

By dealing with homogeneous or heterogeneous Big Data, we consider how different kinds of profiling, such as meta-profiles, which vary depending on the ongoing sequences of Big Data, allow for the establishment of dominions, compatibilities, and incubators of properties that are figured out by the researcher.

We mention, among a large variety of possible searches, the *backward search*, which identifies the paths that led to the current state ex post and allows for learning and making possible generalisations of the paths as profiles for equivalent states.

Non-computable, emergent profiles allow for the tracking of structural dynamics that is invisible to researchers using the dynamics of the same profiles.

The aim of the case of mesoscopic meta-profiling [68] introduced above is to avoid the classical assumption that the detection of any possible properties should be anticipated by analytical, correlational, and statistical properties. According to the conceptual constructivist and De Finetti's way of thinking, the properties should be hypothesised by the researcher first or realised ex post. However, the occurrence of analytical, correlational properties is used as the abductive material for the creativity of the researcher by identifying variables, scales, network representations, and properties.

Here, we consider that *mesoscopic forecasting and understanding* have a dynamical trade-off between their microscopic and macroscopic aspects. This level of representation is considered suitable to represent the *tendency* and *ongoing properties* more-so than the *impreciseness of classical fuzziness*.

The proposal here is to focus on the mesoscopic instead of, or combined with, the fuzziness used in several approaches, such as that used for search engines on the web.

Some generic examples of the *kinds* of properties to look for in the structural dynamics of Big Data which the mesoscopic approaches may facilitate the emergence of are as follows:

- Temporal quantitative trends of aggregations;
- Singularities and their possible recurrence;
- Possible equivalent properties that are necessary or preliminary to the phenomena of emergence;
- Temporally evolving properties.

Social examples of the three kinds of meta-profile considered above may relate to the representation of the maintenance of peaceful conditions, the emergence of conflictive situations, and social trends.

Another area of interest is *cryptocurrency* [69]. A cryptocurrency is a virtual, digital, and decentralised currency whose implementation is based on the principles of cryptography, both for the generation of money itself and for validating transactions.

The implementations use the *peer-to-peer* technologies of peer (non-hierarchical) nodes on networks whose nodes are computers of users scattered everywhere.

Cryptocurrencies have decentralised control; they work through blockchains, a public database of transactions that can be understood as a distributed ledger in which a set of subjects shares IT resources, such as memory, processing, and telecommunication band, in order to make a generally public virtual database that is available to the user community and in which each participant has a copy of the data.

Transactions take place collectively on the network, so there is no "centralised" type of management. In fact, there is no central authority that controls them. Cryptocurrencies are then data and transaction generators that *plump* Big Data [70,71].

Several cryptocurrencies were conceived to induce or facilitate the introduction of subsequent new units of currency in order to place a limit to the amount of money in circulation. This is both to *imitate* the scarcity (with a corresponding effect on the value) of precious metals and to avoid a hyperinflation phenomenon after which users have to use another currency. We mention that Bitcoin, introduced in 2009, was the first decentralized cryptocurrency.

Furthermore, such profiling allows for some peculiar approaches, such as the following:

- The design and detection of *impossible* profiles, for example, artificial profiles to hide the real ones (this may be the case to detect criminal behaviours);
- The consideration of tentatively evolving profiles, for example, to be supported or prevented;
- The exploration of admissibility and compatibility of profiles;
- The identification of incompatibilities and inconsistencies (to be eventually used to disable the creation of a process);
- The identification of equivalences among profiles;
- The identification of composition criteria among profiles;
- The consideration of *evolutionary* criteria (admissible, possible, compatible) with different levels of sensitivity to the initial conditions;
- The consideration of the robustness of profiles.

We mention how this approach may also be considered in linguistics, for instance, it can be related to the Latent Semantic Analysis [72].

We conclude this section by stressing that we do not look for past profiles to explain the current ones, but rather, we look for ongoing profiles (it is their emergence and the properties of their emergence that trace the trajectories). We focus on *profiling* and *querying* Big Data rather than trying to *definitively understand* them and assume that we can then foresee them. *Querying is intended as an ongoing understanding*.

8. Further Research

The methodology outlined here has the potential to be implemented in a variety of approaches in relation to the field of investigation and the interests of the researcher. As a result, several software tools will have to be developed to complement, for instance, those of clustering and cluster analysis that are already in use.

A further line of research may consider the compatibilities, equivalences among fuzzy profiles, semantic fuzzy profiles, and their properties during networking by considering *variable idempotence* (when $M \approx c + kM^2$ as in [73]), as well as other matrix properties *M(t)* that represent the networks and random matrices that allow for possible mathematics-specific descriptions.

Furthermore, it is possible to consider approaches such as networking for Big Data [74,75] and, by using computational network-based approaches, to determine the properties of networked data ([8], pp. 287–304; 64).

Another line of research to be mentioned is related to Quantum Fields Theory (QFT), in which some concepts and approaches may be, at least, inspiring, such as the use of fields as autonomous entities; the existence of non-equivalent representations; and entanglement, where a system cannot be described individually, but only as a superposition of several systems and matrixes $M_s(t)$). As a consequence, a measurement or an intervention on a system instantly determines the variations in the others. This also occurs in the case where systems considered to be overlapping or correlated are spatially or temporally distant.

Moreover, "... the plasticity itself of QFT conceptual structures does not preclude the unexpected occurrence of new models and new achievements helping to better understand what is emergence" ([8], p. 248).

9. Conclusions

In this article, we discussed and considered some new, unusual conceptual frameworks and approaches.

The fundamental new understanding relates to a move from considering Big Data as data sets *within* properties towards considering Big Data as multiple processes to be considered in constructivist ways, for example, by ongoing profiling rather than modelling and understanding rather than foreseeing. Querying Big Data rather than trying to definitively model properties when querying is intended to allow ongoing understanding. We considered the mesoscopic level used to model collective behaviours, such as collective beings, to be the level of representation that is suitable to deal with the ongoing multiplicity of structural dynamics in contrast to dynamics alone.

In particular, we proposed the use of the properties of mesoscopic variables to profile Big Data from a constructivist view, where, as in De Finetti's view of probability, the observer semantically introduces variables and properties. We stated that it is a matter of using mesoscopic profiling that is suitable to detect properties within the context of structural dynamics, where the dynamics relate to the changes of structures, rather than the changes of values regarding the *same* structure. The focus is then on ongoing profiles as they emerge (the properties of the becoming that trace the trajectories). We focused on *profiling* Big Data rather than trying to *definitively understand* them and assuming we can then foresee them (from the past). This is a matter of having *an ongoing mesoscopic understanding* that overlaps with *profiling*. We mentioned the case of non-computable profiles, such as when dealing with natural computing, machine learning, and ecosystems.

Following the proposed interpretations and the possible future research mentioned, it is possible to consider consequential innovative approaches that could be transformed into methodological and technological implementations with fresh ways of dealing with Big Data. In particular, the purpose is to allow the implementation and engineering of tools that can be used to understand and design, rather than to foresee and manipulate. The availability and usage of such tools may facilitate, induce, and support processes of development that are intended to represent multiple, strategic usages of processes of growth.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

References

1. IBM. What Is Big Data? Available online: http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html.May2011 (accessed on 21 May 2018).
2. Minati, G.; Licata, I. Emergence as Mesoscopic Coherence. *Systems* **2013**, *1*, 50–65. Available online: http://www.mdpi.com/2079-8954/1/4/50 (accessed on 21 May 2018).
3. Todde, V.; Giuliani, A. Big Data. A briefing. *Annali dell'Istituto Superiore di Sanità* **2018**, *54*, 174–175. [PubMed]
4. Aggarwal, C.C. *Data Mining: The Textbook*; Springer: New York, NY, USA, 2015.
5. The Vision of IBM: Big Data Is Broken into Four Dimensions: Volume, Variety, Velocity, and Veracity. Available online: https://www.ibmbigdatahub.com/tag/587 (accessed on 5 February 2019).
6. Minati, G.; Pessa, E. *Collective Beings*; Springer: New York, NY, USA, 2006.
7. Minati, G. The non-systemic usages of systems as reductionism. Quasi-systems and Quasi-Systemics. *Systems* **2018**, *6*, 28. [CrossRef]
8. Minati, G.; Pessa, E. *From Collective Beings to Quasi-Systems*; Springer: New York, NY, USA, 2018.
9. Minati, G.; Licata, I. Meta-Structural properties in Collective Behaviours. *Int. J. General Syst.* **2012**, *41*, 289–311. [CrossRef]
10. Gash, H. Constructing constructivism. *Constr. Found.* **2014**, *9*, 302–327.

11. Minati, G. Knowledge to Manage the Knowledge Society: The Concept of Theoretical Incompleteness. *Systems* **2016**, *4*, 26. [CrossRef]
12. Licata, I. Logical openness in cognitive models. *Epistemologia* **2008**, *31*, 177–191.
13. Minati, G.; Penna, M.P.; Pessa, E. Thermodynamic and Logical Openness in General Systems. *Syst. Res. Behav. Sci.* **1998**, *15*, 131–145. [CrossRef]
14. Galavotti, M.C. (Ed.) *Bruno de Finetti Radical Probabilist*; College Publications: London, UK, 2008.
15. Rasetti, M.; Merelli, E. The topological field theory of data: A program towards a novel strategy for data mining through data language. *J. Phys. Conf. Ser.* **2015**, *626*, 012005. [CrossRef]
16. Davenport, T.H. *Big Data at Work*; Harvard Business Review Press: Boston, MA, USA, 2014.
17. Franks, B. *Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics*; Wiley: Hoboken, NJ, USA, 2012.
18. Calude, C.S.; Longo, G. The deluge of spurious correlations in big data. *Found. Sci.* **2016**, *22*, 595–612. [CrossRef]
19. Casacuberta, D.; Vallverdù, J. E-science and the data deluge. *Philos. Psychol.* **2014**, *27*, 126–140. [CrossRef]
20. Nural, M.; Cotterell, M.E.; Miller, J. Using Semantics in Predictive Big Data Analytics. In Proceedings of the 2015 IEEE International Congress on Big Data, BigData Congress, Santa Clara, CA, USA, 29 October–1 November 2015; pp. 254–261.
21. Anderson, C. The end of theory: The data deluge makes the scientific method obsolete. *Wired Mag.* **2008**, *16*. Available online: https://www.wired.com/2008/06/pb-theory/ (accessed on 21 May 2018).
22. Minati, G. Does Systemics still need theories? Theory-less knowledge? In *Systemics of Incompleteness and Quasi-Systems*; Minati, G., Abram, M., Pessa, E., Eds.; Springer: New York, NY, USA, in publication.
23. Coveney, P.V.; Dougherty, E.R.; Highfield, R.R. Big data need big theory too. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2016**, *280*, 1–11. [CrossRef] [PubMed]
24. Cecconi, F.; Cencini, M.; Falcioni, M.; Vulpiani, A. The prediction of future from the past: An old problem from a modern perspective. *Am. J. Phys.* **2012**, *80*, 1001–1008. [CrossRef]
25. Hosni, H.; Vulpiani, A. Forecasting in Light of Big Data. *Philos. Technol.* **2018**, *31*, 557–569. Available online: https://www.researchgate.net/publication/317284370_Forecasting_in_Light_of_Big_Data (accessed on 21 May 2018). [CrossRef]
26. Minati, G. Multiple Systems, Collective Beings, and the Dynamic Usage of Models. *Systemist* **2006**, *28*, 200–211.
27. Ballerini, M.; Cabibbo, N.; Candelier, R.; Cavagna, A.; Cisbani, E.; Giardina, I.; Lecomte, V.; Orlandi, A.; Parisi, G.; Procaccini, A.; et al. Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study. *PNAS* **2007**, *105*, 1232–1237. [CrossRef] [PubMed]
28. De Wolf, T.; Holvoet, T. Emergence Versus Self Organisation: Different Concepts but Promising when Combined. In *Engineering Self-Organising Systems: Methodologies and Applications*; Brueckner, S.A., Di Marzo Serugendo, G., Karageorgos, A., Eds.; Springer: New York, NY, USA, 2005; pp. 1–15.
29. Sawyer, R.K. *Social Emergence: Societies as Complex Systems*; Cambridge University Press: Cambridge, UK, 2005.
30. Licata, I.; Minati, G. Emergence, Computation and the Freedom Degree Loss Information Principle in Complex Systems. *Found. Sci.* **2016**, *21*, 1–19. [CrossRef]
31. Tierny, J. *Topological Data Analysis for Scientific Visualization*; Springer: New York, NY, USA, 2017.
32. Turing, A.M. Computing Machines and Intelligence. *Mind* **1950**, *LIX*, 433–460. Available online: https://www.csee.umbc.edu/courses/471/papers/turing.pdf (accessed on 21 May 2018).
33. Minati, G.; Penna, M.P.; Pessa, E. Towards a general theory of logically open systems. In *Proceedings of the 3rd Systems Science European Congress*; Pessa, E., Penna, M.P., Montesanto, A., Eds.; Kappa: Rome, Italy, 1996; pp. 957–960.
34. Minati, G. Phenomenological structural dynamics of emergence: An overview of how emergence emerges. In *The Systemic Turn in Human and Natural Sciences. A Rock in The Pond*; Urbani, L., Ed.; Springer: New York, NY, USA, 2019; pp. 1–39.
35. Hu, J.; Jin, F.; Zhang, G.; Wang, J.; Yang, Y. A User Profile Modeling Method Based on Word2Vec. In Proceedings of the 2017 IEEE International Conference on Software Quality Reliability and Security Companion (QRS-C), Prague, Czech Republic, 25–29 July 2017; pp. 410–414.
36. Hildebrandt, M.; Gutwirth, S. (Eds.) *Profiling the European Citizen: Cross-disciplinary Perspectives*; Springer: New York, NY, USA, 2008.

37. Ghosh, R.; Dekhil, M. Discovering User Profiles. 2009. Available online: http://ra.ethz.ch/CDstore/www2009/proc/docs/p1233.pdf (accessed on 21 May 2018).
38. Kanoje, S.; Girase, S.; Mukhopadhyay, D. User Profiling Trends, Techniques and Applications. *Int. J. Adv. Found. Res. Comput. (IJAFRC)* **2014**, *1*, 119–125. Available online: https://arxiv.org/ftp/arxiv/papers/1503/1503.07474.pdf (accessed on 21 May 2018).
39. Wassermann, B.; Zimmermann, G. User Profile Matching: A Statistical Approach. 2011. Available online: http://www.thinkmind.org/index.php?view=article&articleid=centric_2011_3_10_30042 (accessed on 21 May 2018).
40. Peirce, C.S. Harvard Lectures on Pragmatism. In *The Essential Peirce: Selected Philosophical Writings, 1893-1913*; Houser, N., Eller, J.R., Lewis, A.C., De Tienne, A., Clark, C.L., Davis, D.B., Eds.; Indiana University Press: Bloomington, IN, USA, 1998; pp. 133–241.
41. Von Foerster, H. Notes pour une epistemologie des objets vivants. In *L'unite de Vhomme: Invariants biologiques et universaux culturels*; Morin, E., Piattelli-Palmerini, M., Eds.; Seuil: Paris, France, 1974; pp. 139–155.
42. Licata, I. Seeing by models: Vision as adaptive epistemology. In *Methods, Models, Simulations and Approaches towards a General Theory of Change*; Minati, G., Abram, M., Pessa, E., Eds.; World Scientific: Singapore, 2012; pp. 385–400.
43. De Finetti, B. *Theory of Probability—A Critical Introductory Treatment*; Wiley & Sons: London, UK, 1975.
44. Aihara, K.; Imura, J.; Ueta, T. (Eds.) *Analysis and Control of Complex Dynamical Systems: Robust Bifurcation, Dynamic Attractors, and Network Complexity*; Springer: New York, NY, USA, 2015.
45. Nicosia, V.; Bianconi, G.; Latora, V.; Barthelemy, M. Growing multiplex networks. *Phys. Rev. Lett.* **2013**, *111*, 058701. [CrossRef] [PubMed]
46. Gillies, D. *Philosophical Theories of Probability*; Routledge: London, UK, 2000.
47. Hair, J.F., Jr.; Black, W.C. *Multivariate Data Analysis*; Pearson: Harlow, UK, 2013.
48. Wu, J. *Advances in K-means Clustering: A Data Mining Thinking*; Springer-Verlag: Berlin, Germany, 2012.
49. Charrad, M.; Ghazzali, N.; Boiteau, V.; Niknafs, A. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *J. Stat. Softw.* **2014**, *61*, 1–36. [CrossRef]
50. Tibshirani, R.; Walther, G.; Hastie, T. Estimating the number of data clusters via the Gap statistic. *J. R. Stat. Soc. B* **2001**, *63*, 411–423. Available online: http://web.stanford.edu/~{}hastie/Papers/gap.pdf (accessed on 21 May 2018). [CrossRef]
51. Miyamoto, S.; Ichihashi, H.; Honda, K. *Algorithms for Fuzzy Clustering: Methods in C-Means Clustering with Applications*; Springer: New York, NY, USA, 2008.
52. Bajec, I.L.; Zimic, N.; Mraz, M. Simulating flocks on the wing: The fuzzy approach. *J. Theor. Biol.* **2005**, *2*, 199–220. [CrossRef] [PubMed]
53. Tettamanzi, A.; Tomassini, M. *Soft Computing: Integrating Evolutionary, Neural, and Fuzzy Systems*; Springer: Berlin, Germany, 2010.
54. Klir, G.J.; Yuan, B. *Fuzzy Sets and Fuzzy Logic: Theory and Applications*; Prentice Hall: Englewood Cliffs, NJ, USA, 1995.
55. Zadeh, L.A.; Klir, G.J.; Yuan, B. (Eds.) *Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems: Selected Papers by Lotfi A. Zadeh*; World Scientific: Singapore, 1996.
56. Minati, L. Remote synchronization of amplitudes across an experimental ring of non-linear oscillators. *Chaos* **2015**, *25*, 123107–123112. [CrossRef] [PubMed]
57. Minati, G.; Licata, I.; Pessa, E. Meta-Structures: The Search of Coherence in Collective Behaviours (without Physics). In Proceedings of the Wivace 2013—Italian Workshop on Artificial Life and Evolutionary Computation (Wivace 2013), Milan, Italy, 1–2 July 2013; Graudenzi, A., Caravagna, G., Mauri, M., Antoniotti, M., Eds.; pp. 35–42. Available online: http://rvg.web.cse.unsw.edu.au/eptcs/paper.cgi?Wivace2013.6 (accessed on 21 May 2018).
58. Alhabashneh, O.; Iqbal, R.; Doctor, F.; Amin, S. Adaptive information retrieval system based on fuzzy profiling. In Proceedings of the 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Istanbul, Turkey, 2–5 August 2015; pp. 1–8.
59. Copeland, B.J.; Posy, C.J.; Shagrir, O. *Computability: Turing, Gòdel, Church, and beyond*; MIT Press: Cambridge, MA, USA, 2013.
60. Mac Lennan, B.J. Natural computation and non-Turing models of computation. *Theor. Comput. Sci.* **2004**, *317*, 115–145. [CrossRef]

61. Yu, S.; Liu, M.; Dou, W.; Liu, X.; Zhou, S. Networking for big data: A Survey. *IEEE Commun. Surv. Tutor.* **2017**, *19*, 531–549.
62. Wu, Y.; Hu, F. *Big Data and Computational Intelligence in Networking*; CRC Press: Boca Raton, FL, USA, 2017.
63. Petri, G.; Scolamiero, M.; Doanato, I.; Vaccarino, F. Topological strata of weighted complex networks. *PLoS ONE* **2013**, *8*, e66506. [CrossRef] [PubMed]
64. Merelli, E.; Pettini, M.; Rasetti, M. Topology driven modeling: The IS metaphor. *Nat. Comput.* **2015**, *14*, 421–430. [CrossRef] [PubMed]
65. Merelli, E.; Rucco, M. Topological characterization of complex systems: Using persistent entropy. *Entropy* **2015**, *17*, 6872–6892. [CrossRef]
66. Becker, K. The Power of Classification: Culture, Context, Command, Control, Communications, Computing. In *Deep Search: The Politics of Search beyond Google*; Becker, K., Stalder, F., Eds.; Studienverlag: Munich, Germany, 2009.
67. Sullivan, D. Search Engine Land. 2017. Available online: https://searchengineland.com/google-now-handles-2-999-trillion-searches-per-year-250247 (accessed on 21 May 2018).
68. Devey, M.; Côté, M.C. The Development and Use of Metadata Application Profiles. The Government of Canada Experience. *Ser. Libr.* **2006**, *51*, 103–115. [CrossRef]
69. Narayanan, A.; Bonneau, J.; Felten, E.; Miller, A. *Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction*; Princeton University Press: Princeton, NJ, USA, 2016.
70. Dean, J. *Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners*; Wiley: Hoboken, NJ, USA, 2014.
71. How Big Data and Bitcoin Go Hand-in-Hand. Available online: http://analyticscenter.com/how-big-data-and-bitcoin-go-hand-in-hand (accessed on 21 May 2018).
72. Dumais, S.T. Latent Semantic Analysis. *Annu. Rev. Inf. Sci. Technol.* **2005**, *38*, 188–230. [CrossRef]
73. Minati, L.; Winkel, J.; Bifone, A.; Oświęcimka, P.; Jovicich, J. Self-similarity and quasi-idempotence in neural networks and related dynamical systems. *Chaos* **2017**, *27*, 043115-1–043115-15. [CrossRef] [PubMed]
74. Srinivasa, K.G.; Siddesh, G.M.; Srinidhi, H. *Network Data Analytics*; Springer: New York, NY, USA, 2018.
75. Bifet, A.; Gavaldà, R.; Holmes, G.; Pfahringer, B. *Machine Learning for Data Streams*; MIT Press: Cambridge, MA, USA, 2018.