Big Data and Biological Knowledge

Maël Montévil, Giuseppe Longo

Big Data and Biological Knowledge[1][2]

Maël Montévil[3] and Giuseppe Longo[4]

Summary

Some authors assert that the analysis of huge databases could replace the scientific method. On the contrary, we argue that the best way to make these new technologies bear fruits is to frame them with theories concerning the phenomena of interest. Such theories hint to the observable that should be taken into account and the mathematical structures that may link them. In biology, we argue that the community urgently needs an overarching theory of organisms that would provide a precise framework to understand lifecycles. Among other benefits, such a theory should make explicit what we can and cannot predict *in principle*.

1. Introduction

Biology is a domain where variation has a fundamental theoretical role. Biological variation is profound and qualitative, and we have defended elsewhere the idea that variation justifies that biology requires its own epistemology. Notably, this variation is the basis of the historicity and the contextual nature of living things (G. Longo and M. Montévil 2014; M. Montévil *et al*. 2016) and it is at the core of the adaptivity and diversity of life. Variation is, in part, due to random phenomena at different levels of organization, and to the many forms of interaction between these levels (bio-resonance, see M. Buiatti and G. Longo 2013), yet it is always canalized by constraints and contexts and may be induced by the context to an extent (M-J. West-Eberhard 2003; Montévil *et al*. 2016; G. Longo 2017). In our perspective, variation is thus an integral component, but not

1 To appear as M. Montévil & G. Longo. Big Data and biological knowledge. In: *Predictability and the Unpredictable: Life, Evolution and Behaviour.* Editors: David Ceccarelli & Giulia Frezza.
2 Extensively revised version of G. Longo & M. Montévil. "Big Data et connaissance biologique" in *Sciences de la vie, sciences de l'information*. Eds. T. Gaudin, D. Lacroix, M.-C. Maurel et al. Paris: ISTE-Editions. 2017.
3 Chaire de recherche contributive, IRI, Centre Pompidou, Paris, France and IHPST, CNRS and Université Paris I, Paris.
Mail: mael.montevil@gmail.com
Homepage: http://montevil.theobio.org
4 Centre Cavaillès, République des Savoirs, CNRS, Collège de France et Ecole Normale Supérieure, Paris, and Department of Integrative Physiology and Pathobiology, Tufts University School of Medicine, Boston.
Mail: giuseppe.longo@ens.fr
Homepage: http://www.di.ens.fr/users/longo/

the only component, of diversity and adaptation, both in phylogenesis and ontogenesis, up to having a crucial role in the etiology of cancer (A. SOTO, G. LONGO and D. NOBLE, 2016). It finally leads to a peculiar form of unpredictability, proper to biological dynamics, since variation is largely based on random phenomena, at all levels of organization (M. BUIATTI and G. LONGO, 2013). As for this issue, note that randomness is not an absolute notion, but it means "unpredictability w.r. to the intended theory" (C. CALUDE and G. LONGO, 2016b). And biological randomness deserves its proper treatment as related to the changing phase space (the pertinent observables and parameters or the space of all possible dynamics) and to the role of rare events, in particular along evolution (MONTÉVIL *et al*. 2016; G. LONGO, 2016).

Biologists are thus confronted with the evolutionary diversity and adaptivity of the living. Moreover, organisms possess an internal heterogeneity which corresponds to their different organs (and organites, in the case of cells): "correlated variations" in the terms used by Darwin, depends both on the internal coherence of each organisms and on the changing eco-systemic conditions.. Faced with these two dimension of biological complexity, the human mind sometimes seems disarmed. In this context, the contemporary possibility of developing immense digital databases in collaborative frameworks is regarded as a major opportunity. But this opportunity is not without peril – and analyses lacking biological meaning is not the least of these perils.

All fields of biological sciences are not equally equipped to use these growing databases. Some fields build on robust theoretical thinking. For example, phylogenetic analyses rely on the conceptual framework of the theory of evolution, extensively enriched in the XXth century. This theory frames the production of knowledge on the basis of data by relying on non-trivial theoretical structures, in particular Darwin's principles ("descent with modification" and "selection"). By contrast, there is no well-established, unified theory to understand organisms, their physiology and their development, in spite of recent advances (see A. MINELLI and T. PRADEU, 2014; A. SOTO, G. LONGO, and D. NOBLE, 2016). Despite decades of informal use, the traditional notion of a genetic program has never acquired a real theoretical status, for a lack of both biological pertinence and of referenrence to a rigorous actual scientific notions (PA. MIQUEL; G. LONGO, 2012). This lasting tradition leads to a causal priority assigned to the molecular level, a priority that is embodied in the nature of the data obtained by high throughput techniques. By contrast, many relevant quantities are neglected by the use of Big Data in biology. For example, the modeling of an organ like the heart requires taking simultaneously into account several levels of organization (D. NOBLE 2006). Similarly, many physicists and biologists emphasize the importance of physical quantities in the determination of biological phenomena. Here physical quantities

refer informally to the forces and fields of classical mechanics. For example, the stiffness of a tissue or the forces exerted by cells are fundamental determinant of a tissue. However, these quantities are not associated with high throughput experimental methods. For example, the interplay of forces in a morphogenetic dynamics is neither measured in genomics, nor proteomics or metabolomics. As a result, we can see that the choice of a theoretical framework impacts directly the quantities that should be measured and analyzed.

Beyond the choice of the quantities relevant to understand a given phenomenon, theoretical frameworks also matter for the analysis of data. Statistical analyses are based on mathematical hypotheses that, in general, correspond to theoretical hypotheses, albeit the latter are sometimes informal or even implicit. The capacity of databases to contribute to the comprehension of phenomena depends on the theoretical view that frames the use of these data and confers meaning to them, as well as on the pertinence of these data in relation to a theoretical frame. In short, there is always a choice, sometimes considered to be "obvious" if not unique, of observables to be measured, of a metric, of criteria of numerical approximation: this choice needs to be made and explicitly so.

The application of Big Data to cancer, for example, is developed in a particular theoretical frame, the somatic mutation theory, where the process of carcinogenesis is conceived as the appearance of cancerous cells by the accumulation of somatic, genetic mutations: "The story of cancer is a story of how the body's complex coding systems go awry through the creation of self-perpetuating errors in cellular replication and growth" (A.R. SHAIKH *et al*. 2014). However, this theoretical point of view encounters major conceptual and empirical difficulties. These difficulties manifest themselves in translational researches and explain the limited medical outcomes of cancer biology despite significant investments. For example, changes in the proportion of deaths due to cancer are not large except in cases which can be interpreted in terms of prevention (R.L. SIEGEL, K.D MILLER and A. JEMAL 2015). One of the most influential advocates of the somatic mutation theory of carcinogenesis acknowledges the difficulties of this genocentric approach and stresses that we are once again faced with the "endless complexity" of these phenomena (R.A. WEINBERG 2014).

Several scholars analyze the situation as the manifestation of a theoretical problem and propose alternative viewpoints about the nature of carcinogenesis (C. SONNENSCHEIN and A.M. SOTO 1999; 2011; S.G. BAKER 2011). These theoretical viewpoints also come with different research strategies, consider different levels of organization and relevant quantities (BERTOLASO M. 2016). However, most of the community stick to the somatic mutation theory. From their perspective, it is then appealing to consider Big Data analysis as a solution permitting the treatment of cancer

while keeping the focus on molecular and more specifically genomic data. This technological solution is called personalized medicine or precision medicine. Precision oncology is advocated by groups such as the Personalized Medicine Coalition and is supported by the US government through the Precision Medecine Initiative program.

More generally, the absence of a theoretical framework for organisms makes particularly seductive a certain rhetoric that goes beyond – if not against – the rational use of data. The omnipotence and autonomy of database analysis is at the center of a contemporary myth. For a decade, several successful articles, including one by Chris Anderson (2008), maintain that the figures speak for themselves: "We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot ... Correlation supersedes causation, and science can advance even without coherent models, unified theories ... No semantic or causal analysis is required." The idea is that "data miners" are capable of detecting correlations and orienting decisions without having to perform any theoretical discussion. So it is no longer a matter of enriching the "obsolete" scientific method but instead of replacing it, in particular by bypassing theoretical thinking. This point of view is associated with the slogan that the larger the database, the easier it is to find relations on the basis of which to act.

2. Immense databases, prediction, and chance

The rhetoric that defend the replacement of the scientific method by the analysis of big databases can be assessed by the use of Mathematics. Theorems enable us to demonstrate the limits of these purely algorithmic methods by showing the impossibility of replacing the scientific quest for meaning by pure "data mining". Theorems at the crossroads of ergodic theory and Ramsey's Theory, a combinatory theory of numbers born in the 1920s and well-developed since then, permit to contradict this use of Big Data (C. CALUDE and G. LONGO 2016; H. HOSNI and A. VULPIANI 2017).

2.1. The deluge of spurious correlations

Let us first consider "Ramsey-type" theorems, used in CALUDE and LONGO (2016). These theorems show that for any correlation between numbers in a database, there exists a number (let us say m) such that any database having at least m elements contains the demanded correlation. Therefore, it is just a matter of size, and it is possible to compute a threshold beyond which all databases (sequence of numbers) will contain a regularity with the stipulated characteristics. In other words, be as precise as you wish about the criterion for correlating pairs, triplets, etc., as well as the minimal number of times that you want to observe them, in what space and over what duration and the manner in which you will divide up your database

(for example, by correlating proximate values, even iterated … according to the preferred criterion). Then the theorems mentioned will tell you how many data to gather in order to achieve those criteria, that is to find some correlation realizing them. More precisely, a regularity in an ensemble of numbers may be established by fixing three parameters, or even more ("arity" of the relation, cardinality of the threshold of interest – how many you wish to have, and the partition of the database...). On the basis of these parameters, we can then calculate a number m, such that any ensemble of numbers A that contains at least m elements will satisfy the required regularity.

We should observe that A is any ensemble and that the only requirement is that A must be "sufficiently large", enormous in fact, since m is growing very rapidly as a function of the given parameters. But being arbitrary, A may be engendered by … dice throwing, measurements of an electron's spin-up/spin-down, a random quantum phenomenon, or random phenomena of any kind (physical, biological, social)… The bigger the database the better, the credulous propagandists of Big Data tell us. Is this number m too big to be encountered in our Universe for a correlation between a sufficient number of elements? Then not all sets of numbers of a cardinality below the Ramsey threshold need to contain the pre-given regularity, yet … lots of them will.

In summary, these results tell us that *any* A that is sufficiently big contains arbitrary, thus potentially "spurious" correlations; moreover, if we ask merely that these correlations appear in a high percentage, but lower than 100% of the ensembles, that is "only" in a reasonably high percentage of ensembles, then we would obtain an m attainable by our databases. In short, this hazard in the huge quantities of numbers is by no means rare. Let us explain.

A finite ensemble of numbers may be considered (algorithmically) "random" when it cannot be engendered by a program smaller than its number of elements. This is a notion of "incompressibility" for sequences of numbers, that may be extended to matrices or other organizations of data in finite dimension. It does not correspond exactly to randomness, yet it is a good "symptom" of randomness: that is an incompressible sequence has a high chance to be random; moreover, asymptotically (for sequences tending to infinite length), it does yield a robust notion of randomness for infinite sequences (C. Calude and G. Longo 2016b). Now, the percentage of ensembles of random numbers in this weak sense (incompressibility) tends toward 100 % (measure 1, to be more precise) when their cardinality grows toward infinity. Infinity is big, even for "data miners" who are the richest in data, yet as soon as we are dealing with ensembles of numbers that are expressed with 2000 bits, for example – which is not out of reach – we approach 80 % of incompressible ensembles (Calude and Longo

2016). So good luck making any kind of use in terms of prediction or action of data that may derive from chance! In every case where chance dominates, it is out of the question for the regularities found by clever data-exploration programs to be of any help at predicting if not acting, precisely because they are the fruit of chance, and they, therefore, may not be reproduced in time and in space, or derived from any causal relation. Thus, it is due to chance that one finds spurious correlations as illustrated in the eponymous book by T. VIGEN (2015, see also the associated website http://www.tylervigen.com/spurious-correlations). Picturesque examples include the correlation between the US spending on science, space and technology which correlates with the suicides by hanging, strangulation and suffocation (r=0.99, from 1999 to 2009) or the number of Japanese passenger cars sold in the US which correlates with the suicides by crashing of motor vehicle (r=0.93, from 1999 to 2009). We leave the causal relevance of these correlations to the reader's appreciation. In (C. Calude and G. Longo 2016), we gave the mathematical arguments that justify these spurious correlations and their high chances to appear.

2.2. Data, prediction and dynamical systems

The analysis of prediction is a central question in meteorology. HOSNI and VULPIANI (2017) present an introductory survey of the problems encountered in this scientific area written by two insiders. The first point is that too many data may kill information and forecasting. The issue was understood by von Neumann and Charney since the 1950s. For example, it follows from the nature of hydrodynamic (and thermodynamics) equations that knowledge and description of data concerning waves of too high or too low frequencies may distort the analysis. So data, possibly implicit in the data bases, concerning nonpertinent phenomena, may incorrectly affect the forecast. Moreover, the larger the database, the larger the physical space required to organize them; that is, the data may belong to description spaces (the spaces of the pertinent observables and parameters) of large or even huge dimension. If the dynamics happens to generate some "attractors" (a precise mathematical notion[5]), then the dimension of the attractors also matters, since the relative unpredictability of future evolutions of the intended dynamical system *grows exponentially* with both the phase space and attractors' dimensions (F. CECCONI *et al.* 2012).

Finally, CECCONI *et al.* (2012) give a further mathematical argument against the abuses of Big Data rhetoric. In linear and non-linear dynamics, in bounded phase spaces, regularities may appear under the form of

5 An attractor describes the asymptotic behavior of a dynamical system, that is to say its behavior after the disappearance of short terms behaviors. For example, the attractor of a dynamics which converge to a single state is this state. More complex situations include limit cycles for dynamics which converge towards an oscillatory behavior and strange attractors in the case of chaotic dynamics.

"recurring phenomena". That is, patterns of the dynamics such as series of observable values that go very close to already traveled paths, may be proven to recur. That is to say they may – and actually will – appear again, a famous theorem by Poincaré, 1892. Yet, as later intuited by Boltzmann and proved by KAC (1947), the recurrence times are immense. If the a-critical Big Data proponents claim that they do have sufficiently large sets of numbers to accommodate recurrence and thus "predict", then they surely fall under the case analyzed in section 2.1. That is, their database must be so huge as to exceed the cardinality limits given by Ramsey theorems, beyond which one finds a "deluge of spurious correlations" in *any* database. The conditions necessary to use Big Data strategies for these dynamics are exactly the ones which lead to the appearance of spurious correlations. As a result, their use for prediction and action is not a valid strategy: a correlation does not need to recur (i.e to continue in time) nor to be due to any "causal" structure – beyond certain large sizes, today accessible to Big Data, they are "meaningless" or due only to the size of the database.

3. A few remarks on biological unpredictability

In the introduction, we hinted to the idea that biological variation plays a fundamental theoretical role in biology. The principle of variation that we have proposed entails that biological objects cannot be defined theoretically like in physics (MONTÉVIL *et al* 2016).

In physics, objects are assumed to follow stable equations which can be found on the basis of quantitative transformations (symmetries) and invariants under these transformations. These transformations define the space of possibilities. Changes are then quantitative changes of state in this predefined state space. By contrast, in biology, we defend the notion that changes also impact these invariants and symmetries (LONGO and MONTÉVIL 2014). As a result, variation is also a variation of the relevant equations and a biological object cannot be defined by its invariants and symmetries. Accordingly, the space of possibilities is not a biological invariant, instead it can change over time. Methodologically, it is not possible to assume the existence of an invariant mathematical structure underlying the biological object of interest and to probe this mathematical structure by experiments.

Nevertheless, there are elements endowed with a restricted stability in biological objects. We call "constraints" these relatively stable elements which play a causal role on the processes that they constrain. Constraints are only stable for a limited time and can only be used as invariants at a given time scale. In an organism, constraints mutually stabilize and reconstruct each other so that the organism can maintain itself over time. With M. Mossio, we call this idea closure of constraints (M. MONTÉVIL AND

M. MOSSIO 2015) and we have proposed the principle of organization which states that closure of constraints is a hallmark of biological organisms (M. MOSSIO, M. MONTÉVIL and G. LONGO 2016). In line with previous work of Rosen, Varela, Kauffman, etc., the principle of organization is a way to understand the mutual dependencies in an organism and to interpret biological functions. A constraint is a part of the closure of an organism when it is maintained by a process under another constraint of the organism and at the same time contributes to maintain at least another constraint of the organism, thus contributing to maintaining the whole and ultimately itself through the whole.

Let us now discuss a few consequences of this framework when considering Big Data approaches. Following the principle of organization, the relations between the parts of the individual is a fundamental notion. Following the principle of variation, the set of relevant constraints and their mutual dependencies may undergo variations. The ubiquity of variations is precisely why we can talk of an individual and not of generic organisms which would all have exactly the same organizations. In this context, data analysis cannot unravel a stable structure that would be instantiated in all the data points corresponding to different individuals. Instead, these different data points correspond to individuals that are different to an extent: the constraints involved and their relations are slightly different for different individuals. Of course, data analysis may still help when focusing on a few constraints that are stable enough among the individuals considered. However, analyzing jointly the organization of many individuals leads to mixing different organization together and leveling down their specificity.

4. Conclusion

The results cited in section 2 are technical: they belong to the combinatory theory of numbers and to the theory of algorithms or involve non-trivial aspects of dynamical systems theory and ergodic theory. The defenders of what we define here as "Big Data without Theory" and of data-mining algorithms without analyses of meaning aim to disregard questions pertaining to theoretical frameworks. Another way to look at their aim is to say that they defend the idea of a generic theoretical framework that would apply in all kinds of empirical contexts without the need of a specific elaboration of meaning, from physics to social sciences.

In this context, recall that the Theory of Computability was invented in the 1930s by Gödel, Church and Turing in order to prove the existence of undecidable propositions and incomputable functions. More particularly, in our case, variants of results of Ramsey's Theory are situated in the difficult space of "what is computable" (the set of decidable propositions

and computable functions), but such that its "computability cannot be proven" within formal number theory. That is, they allow defining functions that are computable but cannot be proven to be computable within the proper Theory of Computability (Arithmetic) (G. LONGO 2011) – one needs to step outside this theory and use infinitary or geometric tools in the proofs. These methods and objects are totally extraneous to effective computability and discrete Data Types. Thus, as a non obvious consequence of these results, even checking that a correlation is spurious is highly undecidable for a machine. Instead, it happens that we can generally detect the spurious correlations as in the examples above, whenever we have reasonably good, meaningful theories of many aspects of the world: one can give good reasons why the relation between the number of Japanese passenger cars sold in the US and the number of suicides by crashing of motor vehicle are spurious, in principle (or if it applies, search for a meaningful correlation...).

Mathematical theories such as computability demonstrate their own limits in the possibilities of computations and prediction by "negative results" that are present at the origin of scientific knowledge and characterize it. Once we have grasped the importance of the limits of the myths that "all is algorithmic" or "all is computable", we may make a better use of these immense quantities of data that computer networks make available, which is a great chance for science in every domain, including biology. Once we clarify the hypotheses that make us choose certain observables and not others, and choose measures suitable to the objectives of the knowledge that we are adopting, then digital information can help conjecture or corroborate a theory or a sketch of it, even produce new understanding. Whether it precedes or is propelled by data analysis, it seems urgent and necessary to develop theoretical frameworks for understanding organisms. In this context, we are engaged in a collaborative and interdisciplinary effort whose latest results are contained in a special issue of *Progress in Biophysics and Molecular Biology: From the century of the genome to the century of the organism: New theoretical approaches* (G. LONGO, A.M. SOTO and D. NOBLE 2016).

5. References

(Most papers by the authors are downloadable from their web pages).

ANDERSON C. 2008. *The end of theory: The data deluge makes the scientific method obsolete*. In «WIRED».

BAKER S.G. 2011. *TOFT better explains experimental results in cancer research than SMT.* In «Bioessays», 33: 919–921. doi: 10.1002/bies.201100124

BAILLY F., LONGO G. 2006 *Mathématiques et sciences de la nature. La singularité physique du vivant*. Hermann, Paris.

BERTOLASO M. 2016. *Philosophy of Cancer. A Dynamic and Relational View*. Springer. Dordrecht. Doi: 10.1007/978-94-024-0865-2

BUIATTI M., LONGO G. 2013. *Randomness and Multi-level Interactions in Biology*. In «Theory in Biosciences», vol. 132, n. 3:139-158. doi: 10.1007/s12064-013-0179-2

CALUDE C., LONGO G. 2016. *The Deluge of Spurious Correlations in Big Data*. In «Foundations of Science», 1-18, March, 2016, 2016. doi: 10.1007/s10699-016-9489-4

CALUDE C., LONGO G. 2016b. *Classical, Quantum and Biological Randomness as Relative Unpredictability*. Invited Paper, special issue of Natural Computing, Volume 15, Issue 2, pp 263–278, Springer. doi: 10.1007/s11047-015-9533-2

CECCONI F., CENCINI M., FALCIONI M. and VULPIANI A. 2012. *Predicting the future from the past: An old problem from a modern perspective*. In «American Journal of Physics», 80, 11, 1001—1008. doi: 10.1119/1.4746070

HOSNI H. and VULPIANI A. 2017. *Forecasting in light of big data*. In «Philosophy & Technology». doi: 10.1007/s13347-017-0265-3.

LONGO G. 2011. *Reflections on Concrete Incompleteness*. In «Philosophia Mathematica», 19(3): 255-280. doi: 10.1093/philmat/nkr016

LONGO G. 2017. *How Future Depends on Past Histories and Rare Events in Systems of Life*. In «Foundations of Science», doi 10.1007/s10699-017-9535-x.

LONGO G. and MONTÉVIL M. 2014. *Perspectives on Organisms: Biological Time, Symmetries and Singularities*. Springer, Berlin. doi: 10.1007/978-3-642-35938-5

MONTÉVIL M. and MOSSIO M. 2015. *Biological organisation as closure of constraints*. In «Journal of Theoretical Biology», 372, 179 – 191. doi: 10.1016/j.jtbi.2015.02.029.

MONTÉVIL M., MOSSIO M., POCHEVILLE A. and LONGO G. 2016. *Theoretical principles for biology: Variation*. In «Progress in Biophysics and Molecular Biology», 122 , 36– 50. doi: 10.1016/j.pbiomolbio.2016.08.005.

MOSSIO M., MONTÉVIL M. and LONGO G. 2016. *Theoretical principles for biology: Organization*. In «Progress in Biophysics and Molecular Biology», 122, 24 – 35. doi: 10.1016/j.pbiomolbio.2016.07.005.

NOBLE D. 2006. *The Music of Life: Biology beyond the Genome*. Oxford University Press, Oxford.

SHAIKH A.R., BUTTE A.J., SCHULLY S.D., DALTON W.S., KHOURY M.J. and HESSE B.W. 2014. *Collaborative Biomedicine in the Age of Big Data: The Case of Cancer.* In «J Med Internet Res» ; 16(4):e101. Doi: 10.2196/jmir.2496

SIEGEL R. L., MILLER K. D. AND JEMAL, A. 2015. Cancer statistics, 2015. In «CA: A Cancer Journal for Clinicians», 65: 5–29. doi:10.3322/caac.21254

SONNENSCHEIN, C. and SOTO A.M. 1999. *The Society of Cells: Cancer and Control of Cell Proliferation.* Springer Verlag, New York.

SONNENSCHEIN C. and SOTO A.M. 2011, *The tissue organization field theory of cancer: a testable replacement for the somatic mutation theory.* In «Bioessays» ; 33(5):332-40. doi: 10.1002/bies.201100025.

SOTO A.M. and LONGO G., (editors). 2016. *From the century of the genome to the century of the organism: New theoretical approaches.* Special issue of «Progress in Biophysics and Molecular Biology», Vol. 122, Issue 1, Elsevier, 2016, doi: 10.1016/j.pbiomolbio.2016.

VIGEN T. 2015. Spurious correlations. Hachette Books.

WEINBERG R.A. 2014, *Coming Full Circle—From Endless Complexity to Simplicity and Back Again*. In «Cell», Volume 157, Issue 1, 27, Pages 267-271, ISSN 0092-8674. doi: 10.1016/j.cell.2014.03.004

WEST-EBERHARD M.-J. 2003, *Developmental plasticity and evolution.* Oxford Univ. Press, New York.