Is big data enough? A reflection on the changing role of mathematics in applications

Domenico Napoletani, Marco Panza, Daniele Struppa

Is big data enough? A reflection on the changing role of mathematics in applications

Domenico Napoletani [*], Marco Panza [†], and Daniele C. Struppa[‡]

The advent of computers, and especially high performance computers, has had a dramatic impact on the way in which mathematics is done, and even more on how mathematics is applied, as demonstrated by the growth of computational mathematics as well as what goes under the name of "experimental mathematics", to which a journal is now devoted. More importantly, computers are now used to perform highly complex computations in order to apply mathematical models to a variety of empirical problems that could never be attacked otherwise. It is in this way that mathematics is often now applied in different branches of biology (think of genomics and proteomics) as well as to social sciences, and in what now goes under the generic term of "big data".

High performance computing has also brought changes in the way we think about mathematics, its power, its methods, and the way in which it can be used to solve problems. For this reason, we would like to do something unusual in a mathematical journal, by bringing the mathematical community into one of the discussions that are taking place among philosophers of mathematics and of science. While many mathematicians are weary of asking broad methodological questions, because the danger of being vague and unclear is all too real, we also think that we must have this discussion concerning the modalities in which mathematics is applied, if we don't want to be trapped by our own assumptions, and miss out on more fruitful approaches to understanding reality. This a discussion about the ways computational mathematics has not only changed our approach to science, but even our way to understand a phenomenon.

[*]Institute for Quantum Studies, Chapman University, Orange, CA, 92866. Email:napoleta@chapman.edu

[†]IHPST, UMR 8590, CNRS, Univ. Paris 1, and ENS Paris. Email: marco.panza@univ-paris1-fr

[‡]Schmid College of Science and Technology, Chapman University, Orange, CA 92866. Email: struppa@chapman.edu

We believe it is possible to identify four methodological motifs that are closely related to each other and whose elucidation may make explicit the approach to problem solving in data analysis and statistical learning. By using short and evocative labels, we could say that these motifs concern, respectively, the *microarray paradigm*, the *preeminence of historical phenomena*, the *conceptualization of developmental processes*, and the *principle of forcing*, which we have introduced and partially expanded upon in [23, 24, 25], also in response to some acute commentaries [14, 20]. Here we would like to simply focus on their rationale and potential significance in clarifying the underlying trends of quantitative, data-driven science.

The first methodological motif, the microarray paradigm, is derived by considerations that have become quite important in modern molecular biology. As it is well known, DNA microarray technology (see for example [2]) allows the scientist to capture and visually represent, on an extremely large array of microscopic sites, information on the expression level of short strands of messenger RNA (mRNA), extracted and amplified from a given cell population. When the large amount of data derived from DNA microarrays is coupled with simple clustering techniques, it is often sufficient not only to differentiate cell populations from distinct tumors, but also to determine the outcome of a given therapy, without the need for an understanding, in each tumor population, of the actual role of each of the mRNA strands whose expression level is quantified by the microarrays.

Thus in [23] we have spoken of the microarray paradigm to indicate how much modern data analysis is supported by the belief that sufficiently large data collected from a phenomenon will allow to answer any question about the phenomenon itself, if treated with appropriate methods, and assisted by powerful enough algorithms. Significant methods in statistical learning theory that embody the microarray paradigm include the aforementioned clustering techniques, in their more traditional hierarchical forms [31], but especially in their more unstructured versions such as the affinity propagation method recently introduced in [11], that exploits a completely local passage of information among data points to obtain their fast and accurate splitting into clusters, whose best representative, the 'exemplar' is also identified. Boosting [10] is another striking technique, perhaps the most transparent embodiment of the microarray paradigm, where a large number of classifiers, slightly better than random guesses, are combined, boosted, to build a new classifier that is much more accurate than its components (cf. [13], chapter 10). Finally, we mention here the recent field of 'nonlinear manifold learning', a collection of methods to find low dimensional objects, preserving some type of local, neighborhood structure, in unstructured, high

dimensional data (see [30, 28] and [15], chapter 16).

As we can see from these examples, within the microarray paradigm, answers are found through a process of automatic fitting of the data to models that do not carry any structural understanding beyond the actual solution of the problem itself, a distinctive lack of knowledge which we wanted to emphasize in [23] by speaking of "agnostic science". The somewhat naive belief in the assumptions of the microarray paradigm was popularized in a recent cover article in *Scientific American* [32], where the author discusses how the collection of unprecedented amount of data could allow the construction of a computing model to predict the future, and even earlier in the opinion piece of Chris Anderson in *Wired* magazine [1] that envisioned a future of atheoretical, automatic science.

Our objective in [23] was less concerned in stating that such paradigm has gained ground, or why this has happened, than in uncovering its structure, and the assumptions it depends upon. Indeed physicists often work and approach their science from a very different point of view. Regardless of how complex nature may be, there is a belief that a fundamental explanation can be found, and that a theoretical, explicative model can reached. So, for example, Newtonian dynamics is considered a fundamentally successful description of the universe, and an archetype of the way science should look like, irrespective of the fact that, within its framework, a solution of a basic question such as the three-body problem is, effectively, not computable. And yet these shortcomings have not led physicists to abandoning Newtonian dynamics: within non-relativistic velocities, and as long as the bodies we study are not too small, Newtonian mechanics is still (rightly) considered a perfectly good model of reality, indeed one of the most successful and enduring.

If it is not the lack of computability (as in the three-body problem) or the lack of precision (as it happens for the relativistic effects that Newtonian dynamics fails to account for or predict), then what does require the agnostic approach embodied in the microarray paradigm? We suggest that such an approach is required when the phenomena we want to deal with are historical in a sense that needs some explanation: at least in an intuitive sense, we call 'historical' those phenomena whose development can only be constrained locally (in time and/or space) by (potentially multiple) optimization processes acting on subsets of variables, and in such a way that the functions to be optimized change over long periods of time. It appears that many of the sciences, which have been less receptive to the classical process of mathematization, concern this sort of phenomena. Biology, economics, the social sciences in general, all are examples of disciplines that study phe-

nomena whose development seems to be historical in the sense we suggested above. Our second broad methodological motif is thus the preeminence of historical phenomena in contemporary science.

Fitness landscapes, introduced by Sewall Wright in [34], best exemplify and motivate the definition of historical phenomena: the evolution along such landscapes is partially shaped by local optimization constraints but, crucially, these constraints change dramatically their nature over time, so that such evolution is not likely to be fully described by any single global optimization process. The selective pressure, due to varying environmental conditions, on the genotypes of a given population is an example where a local search for an optimal phenotype is subject to constantly varying constraints. Recalling that alternate versions of a same gene are called 'alleles', the basic idea of fitness landscapes is to represent a population as a point in a high dimensional space determined by its average allele frequencies, and to represent the average fitness of the population by the value of a corresponding function; the graph of the function generates a multidimensional surface (which is just the fitness landscape) and the potential evolution of the population is described by the local maximization of the fitness on the graph. Note that the landscape itself will slowly change when the evolution of several competing populations (not necessarily from the same species) is considered at once, since the fitness of one of them will affect the environment of the other (see [12] for a recent overview of these issues).

Coming back to the general definition of historical phenomena, we can assume that the time scale at which there is a switching from one local optimization process to another in the development of such phenomena, is much longer than the time scale of the optimization processes themselves. Again, fitness landscapes could provide a motivation for this assumption, since the environmental conditions, and the shape of the landscape, change slowly with respect to the change of the position of the individual genotype points ([12], page 1613). Note also that not all variables are likely to be subject to selective local optimization at the same time.

Now it is possible, indeed it happens often, that specific questions about an historical phenomenon can be reduced to simple models, that require knowledge of only a few key measured quantities. But the models derived to answer each individual question are not only usually data-driven, but often unable to answer other relevant questions on the same phenomenon. And the methods exemplifying the microarray paradigm we mentioned earlier certainly lead to temporary models that have this characteristic. This inability to gain a global understanding is not surprising, if we think about the randomness and relative independence of the selective optimization pro-

cesses that shapes individual characteristics of the current state of an historical phenomenon.

It may be that a full, structured understanding that potentially leads to answers for a wide enough class of questions about a typical historical phenomenon would require knowledge of its entire history, or the development of a proxy, a data-driven model that is likely as complex, and opaque, as the phenomenon itself, a perspective reminiscent of incompressibility ideas in Kolmogorov complexity [21], but in the context of a whole set of suitable potential questions. The methodological danger is that the flood of data generated by our innumerable measuring devices may convince us that data is enough, that there is nothing beyond the microarray paradigm, and that opaque, enormous, data-driven models are the privileged way to approach phenomena, even though they become so similar to the famous map of Borges [4], that was useless, since it was as big as the geography it was supposed to describe.

So the problem arises of how we can gain meaningful understanding of historical phenomena, given the tremendous potential variability of their developmental processes. Indeed, several techniques in mathematical modeling already show the usefulness of a partial historical modeling of key variables, evolving in time, describing a phenomenon, we think here of nonlinear time series analysis methods [17], where modeling of random and/or nonlinear processes includes knowledge of its past states at several time points; or of forecasting methods where ensembles of initial conditions are used to best predict the future state of complex systems, such as the ensemble Kalman filter [16, 9], or the even less structured particle filters method [19]. All these methods show a distinct awareness of the relevance of modeling incidental, historical developmental processes. And we should not forget that many successful heuristic methods for optimization are directly inspired to the idea of fitness landscapes. One example is the field of evolutionary algorithms [8], modeling optimization of a function as a form of reproduction of solutions, that allows tentative solutions to the optimization problem to generate new ones through mutation and genetic crossover, and therefore allowing unexpected changes of their fitness, seen as the evaluation of the function to be optimized at the tentative solution. Another example is particle swarm optimization [18], where the search for optimal solutions to a problem is seen as a collective process in which individual, tentative solutions are changed in time not only by their tendency to improve their current fitness (in the fitness landscape generated by the function to be optimized), but also by their tendency to retain close contact with the best known tentative solution.

Still, all these methods work more at the operational level, where model

classes have already been chosen to solve a certain problem. They do not constitute, taken individually, a coherent, conceptual shift in the way to approach historical phenomena, and their complexity. We can try to imagine more radical ways in which such complexity can be tackled by looking at existing attempts within biology to conceptualize rules of development. In [22] for example a general "principle of biological inertia" is introduced, to give a broad conceptual basis to the multiform expression of default dynamics in developmental biology. Roughly stated, the principle asserts that, without external disturbances or internal (genetic) control, there is a "local self-perpetuation of cell-level dynamics" ([22], page 119). Among the many embodiments of this principle, in the context of embryo development, biological inertia would correspond to stating that embryos have a tendency to reproduce spatially the same basic structure indefinitely. This embodiment of biological inertia could be seen in the context of historical processes as due to local optimization functions that tries to maximize replicates, or spatial distribution, of a basic template. However, in a real system we cannot expect this inertial behavior of developmental systems to be indefinitely exact ([22], page 123), therefore the local optimization process will break down in the timeframe of full development and a complex, inhomogeneous organism may arise.

Note the subtle and nonconventional use of ideas from physics where only analogy is at work: the principle of biological inertia parallels the principle of inertia in mechanics, but only at the very general level of establishing the equivalent for biological systems of a state of rest or dynamical invariance. At the same time, what is essential in the idea of physical inertia is retained, and its usefulness is seen in the power of conceptualizing the developmental process itself, rather than in the ability to identify and predict the final outcome of the process. This logically rigorous and yet informal use of ideas from mathematics and physics to define principles that partially govern biological processes might become the standard for a proper structuring of historical sciences, exactly because individual historical phenomena do not seem amenable to a compact explanation of their structure, and only the development that led to their current state may be open to meaningful theorization, a shift advocated by our third methodological motif, the conceptualization of developmental processes.

Indeed, the theory of evolution could be seen as the archetypical example of this motif. Such theory, in itself, is not mathematical, and does not allow quantitative predictions, but it has its internal logical structure, and provides a conceptual scaffolding that has inspired biology, and specific mathematical models, since its inception. As an aside, we believe that

our point of view is in line with the recent commentary by Wilson [33], who suggested that profound ideas in science need not to be mathematically profound; in this perspective, mathematization is an overflow of the richness and potential of an idea, not a condition of its power.

What is crucial for historical sciences is that the conceptualization of a developmental process may not even lead to a structural, full understanding of the state of the resulting phenomenon. The phenomenon may forever remain hidden to our understanding, and the microarray paradigm will stand as the only way to find quantitative answers to most problems we will ask about it. Conceptualization, at best, can provide the ground, the language, on which to develop data-driven, agnostic methods to solve problems. This acknowledgment does not hinder however the tremendous potential of finding ways to use mathematical structures to solve problems about historical phenomena.

To reflect on this potential, it may be useful to change our focus. Up to now our discourse moved from the microarray paradigm to an attempt to characterize the types of phenomena that most likely will require its application. But we could also reverse this viewpoint, taking for granted the applicability of the microarray paradigm (i.e. that we have enough data), and trying to understand, operationally, how to apply it.

One way to do that is of course to give free rein to statistical learning, data-driven techniques. But here we want to discuss a more general organizational motif: the principle of forcing[1], which we introduced in [23]. In that paper, we suggested that several disparate techniques developed to apply sophisticated mathematics to empirical problems can be brought together under a common methodological viewpoint, the idea of forcing mathematical ideas and methods on the data. More precisely, by forcing we mean the application to a problem of powerful techniques — such as multiscale methods as applied to image processing and numerical solutions of differential equations ([5, 6]), continuity and functional data analysis as applied to regularization and statistical analysis [26, 27], or topological graph analysis as applied to classification of molecular structures [3]— not on the basis of a previous evaluation that these techniques fit with the relevant phenomenon because of its specific nature, but rather on the basis of *a priori* confidence in the power and flexibility of these techniques [23], and using the large amount of available measurements to adapt the technique to the phenomenon. This

[1]We named this principle 'forcing' in [23] because this term is strongly evocative of the general methodological approaches it pertains to. It has however no significant relation to the method introduced by Paul Cohen, now of common use in set theory, that goes by the same name.

is an approach that contrasts with the usual idealization process in modeling, where there is a progressive stripping away of details from the phenomenon, to reach a simplified image of the same, eventually amenable to analytical treatment by a mathematical technique (whose usefulness is often suggested by the idealization process itself).

The example of functional data analysis is perhaps the simplest instance of forcing, and the most telling: a phenomenon may be discrete, and yet, if there is enough data (microarray paradigm) we can force regularization to be able to treat the data as if they were continuous or even smooth, and therefore access the full machinery of analysis. A more recent example of forcing, along the same general trust of functional data analysis, is diffusion geometry [7, 29]. The organizational principle of this theory can be seen, at a very general level, as the belief that, regardless of the whether the data available in the empirical problem are categorical, and/or discrete, it is useful to define a notion of geometrical manifold associated to the data, because in this way the whole apparatus of functional analysis on manifolds can be adapted and used to solve the problem itself. For example, the review [7] describes how scientific journals can be seen as points on a low dimensional manifold, built by first associating each article to a vector of the frequencies of preselected words in the article itself, and then by projecting the cloud of points associated to the whole set of articles onto the directions of maximal variance. We refer to [23, 7] and to the relevant primary literature [29] for more details.

To be sure, forcing may be viewed as a coarse and willful attempt to use mathematics for specific purpose, but are there limits to the applicability of forcing? Is it plausible to think that, for sufficiently large and diverse data sets, any mathematical structure can be forced on them in a computationally efficient way to solve problems?

While we do not have answers to these questions, the weakening of the relation between individual models and phenomena should make us think more deeply about the role of mathematics in science. We suggested above that historical phenomena may not be amenable to uniform, structured reduction to simple models, and in these cases a data analysis approach to each problem instance is most likely the best that can be expected when approaching them, whether by forcing, or by standard statistical learning and classification techniques. And these data-driven applications of mathematics lead us to an intriguing version of the famous question of Wigner about the unreasonable effectiveness of mathematics: how can classification methods, which are essentially function fitting on data, be so successful at predicting phenomena?

Our suggestion is that the four methodological motifs we highlighted may give us a context for asking such questions and for a comprehensive reflection on the ways modern science is so effective at problem solving. We are also encouraged by the interdependence of the motifs: the notion of historical phenomena, and the principle of forcing seem pertinent only in the context of large data sets, and therefore are deeply dependent on the microarray paradigm; and defining historical phenomena naturally shifts the emphasis from phenomena's states to the developmental processes that generate them.

Surely science may have become agnostic, and data analysis methods are often incapable of providing understanding of phenomena, they only give answers to our problems. But this does not imply that its methods should not be amenable to understanding. On the contrary, it is exactly the absence of understanding of phenomena that brings urgency to a widely shared methodological and epistemological reflection of the scientific community. Mathematics may not work necessarily for historical sciences the way it did for physics, but that does not mean that it has to reduce itself to blind computations, and principles such as biological inertia show that it is possible to gain deep insight into the rules of this historical development.

And therefore this is, in the end, an invitation to mathematicians, to approach biology and other historical sciences on their own terms, a process that frustrates superficial knowledge of each field, and challenges us, if we want to be relevant, not so much to be interdisciplinary, as to be scientifically bilingual. We may discover that what is essential in a field, and the true linchpin of its conceptualization, is often very different from what we deem profound or interesting in our own mathematical disciplines.

Acknowledgments

We thank the three anonymous referees, a mathematician, a philosopher, and a biologist, for their thoughtful and very helpful comments.

References

[1] C. Anderson, The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired, vol. 16, n. 7, 2008. Available at: http://www.wired.com/wired/issue/16-07.

[2] P. Baldi and G.W. Hatfield, *DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modeling*, Cambridge University Press, Cambridge, New York, 2002.

[3] D. Bonchev and D. H. Rouvray. *Chemical Graph Theory: Introduction and Fundamentals.* Abacus Press, New York, 1991.

[4] J.L.Borges, Del Rigor en Sciencia, in *Historia Universal de la Infamia*, Emece, Buenos Aires, 1954.

[5] A. Brandt, and O. Livne, O. *Multigrid Techniques: 1984 Guide with Applications to Fluid Dynamics*, Revised Edition, Society for Industrial and Applied Mathematics, Philadelphia, 2011.

[6] A. Brandt, Multiscale Scientic Computation: Review 2001. In Multiscale and Multiresolution Methods: Theory and Applications, T. J. Barth, T. F. Chan, R. Haimes (eds.), Springer Verlag, 2001.

[7] R. R. Coifman and M. Maggioni. Geometry, analysis and signal processing on digital data, emergent structures, and knowledge building. SIAM News, 41(10), 2008.

[8] K. A. De Jong, *Evolutionary Computation*, MIT Press, Cambridge, 2006.

[9] G. Evensen, Advanced Data Assimilation for Strongly Nonlinear Dynamics, Monthly Weather Review, vol. 125, pp. 1342-1354, 1997.

[10] Y. Freund, R. E. Schapire, A decision-theoretic generalization of online learning and an application to boosting. Journal of Computer and System Sciences, 55(1):119–139,1997.

[11] B. J. Frey, D. Duek, Clustering by Passing Messages Between Data Points. Science Vol. 315 no. 5814 pp. 972-976, 2007.

[12] S. Gavrilets, Fitness Landscapes, In *Ecosystems*. Vol. 2 of Encyclopedia of Ecology, edite by S.E.Jørgensen and B.D. Fath, pp. 1612-1615 Oxford: Elsevier, 2008.

[13] T. Hastie, R. Tibshirami, J. Friedman, *The Elements of Statistical Learning*. Springer, New York, 2001.

[14] P. Humphreys, Data Analysis: Models or Techniques? Foundations of Science, August 2013, Volume 18, Issue 3, pp 579-581.

[15] A. J. Izenman, *Modern Multivariate Statistical Techniques*, Springer, New York, London, 2008.

[16] E. Kalnay, *Atmospheric modeling, data assimilation and predictability*, Cambridge University Press, New York, 2003.

[17] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, Cambridge University Press, Cambridge, 2003.

[18] J. Kennedy, R. C. Eberhart, Y. Shi*Swarm Intelligence*. Morgan Kaufmann, San Francisco, 2001.

[19] H. Kotecha and P. M. Djuric, Gaussian particle filtering, IEEE Trans. Sig. Proc., vol. 51, pp. 2592–2601, 2003.

[20] J. Lenhard, Coal to Diamonds, Foundations of Science, August 2013, Volume 18, Issue 3, pp 583-586

[21] M. Li, P. M.B. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, New York, 2008.

[22] A. Minelli, A principle of Developmental Inertia, in *Epigenetics: Linking Genotype and Phenotype in Development and Evolution*. Edited by Benedikt Hallgrimsson and Brian K. Hall. Berkeley, CA: University of California Press, 2011.

[23] D. Napoletani, M. Panza, and D.C. Struppa, Agnostic science. Towards a philosophy of data analysis, Foundations of Science **16** (19), 1–20 (2011).

[24] D. Napoletani, M. Panza, and D.C. Struppa, Artificial diamonds are still diamonds, Foundations of Science, August 2013, Volume 18, Issue 3, pp 591-594.

[25] D. Napoletani, M. Panza, and D.C. Struppa, Processes rather than descriptions?, Foundations of Science, August 2013, Volume 18, Issue 3, pp 587-590.

[26] J.O. Ramsay, B.W. Silverman, *Functional Data Analysis*, Springer, New York, 1997.

[27] J. O. Ramsay and B. W. Silverman. *Applied Functional Data Analysis*. Springer, New York, 2002.

[28] S. T. Roweis, L. K. Saul, Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science, Vol. 290 no. 5500 pp. 2323–2326, 2000.

[29] A. D. Szlam, M. Maggioni, and R. R. Coifman. Regularization on graphs with function-adapted diffusion processes. Journal of Machine Learning Research, 9:1711–1739, 2008.

[30] J. B. Tenenbaum, V. de Silva, J. C. Langford, A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science 290 (5500): 2319–2323, 2000.

[31] J. H. Ward, Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association Volume 58, Issue 301, 1963.

[32] D. Weinberger, The Machine That Would Predict the Future, Scientific American, December 2011.

[33] E.O. Wilson, Great Scientists $\neq$ Good at Math, The Wall Street Journal, April 5, 2013 (also reprinted in Notices A.M.S. **60** (7), 837-838.

[34] Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding, and selection in evolution. Proceedings of the Sixth International Congress on Genetics. pp. 355-366.