The Agnostic Structure of Data Science Methods

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

November 23, 2018

1 Introduction

In this paper we want to discuss the changing role of mathematics in science, as a way to discuss some methodological trends at work in big data science. More specifically, we will show how the role of mathematics has dramatically changed from its more classical approach. Classically, any application of mathematical techniques requires a previous understanding of the phenomena, and of the mutual relations among the relevant data; modern data analysis appeals, instead, to mathematics in order to identify possible invariants uniquely attached to the specific questions we may ask about the phenomena of interest. In other terms, the new paradigm for the application of mathematics does not require any understanding of the phenomenon, but rather relies on mathematics to organize data in such a way as to reveal possible invariants that may or may not provide further understanding of the phenomenon per se, but that nevertheless provide an answer to the relevant question.

However, postponing or giving up altogether the understanding of phenomena and making it dependent on the application of mathematics calls for a different kind of understanding, namely the understanding of the reasons that make the mathematical methods and tools apt to answer a specific question.

A current explanation of the power of data analysis (whose pervasiveness in popular literature dispenses us from making specific references) is that there is, in fact, not much to be understood: mathematics merely consists of a number of disconnected methods and, drowned in sufficiently many and diverse data, succeeds only because of the omnipotence of the data themselves; a new Wignerian paradox of "unreasonable effectiveness"; this time, however, the effectiveness is assigned to data, not to mathematics: a sort of revenge of facts against their mathematization.

We reject such an answer as both immaterial and unsupported. But in our rejection we do not argue (as it is often done) that any exploration of data is doomed to failure if the scientist does not have some previous understanding of the phenomenon (in other words, we do not

[*]University Honors Program and Institute for Quantum Studies, Chapman University, Orange (CA)
[†]CNRS, IHPST (CNRS and Univ. of Paris 1, Panthéon-Sorbonne) & Chapman University, Orange (CA)
[‡]The Donald Bren Presidential Chair in Mathematics, Chapman University, Orange (CA)

oppose data analysis' methods by defending the classical approach). Rather we observe the effectiveness of data analysis in the absence of previous understanding, and wonder about what makes this possible.

The question is far from simple. Much less simple than an ideological exaltation of the power of data or an, equally ideological, rejection of their ability to answer questions, at least when supported by appropriate algorithms. Its difficulty make us unable to even suggest here any exhaustive and/or definitive answer. We will simply observe that no answer is possible unless we engage in a technical inquiry on how these algorithms work, and suggest a largely schematic account of their *modus operandi*. This account strongly relies on the results of our previous works, in particular [9], [10], [11], which we will reorganize in a comprehensive and hopefully coherent way. Furthermore, we will identify a possible direction for future researches, and specifically a promising perspective from which the question can be tackled.

Before we come to this, we need to make an important proviso regarding the notion of understanding. In this paper we will not even try to offer a precise characterization of this notion, as this would take us into a different and more complex direction. Nevertheless, we can still offer some clarification about the way we appeal to it.

In [9] we have discussed the lack of understanding proper to the current big data methods. In so doing, we have avoided to borrow any general and univocal notion of understanding for an empirical (physical, biological, biomedical, social, etc.) phenomenon (which is, by the way, in no way provided by the specialized literature). We simply observed that big data methods typically apply not to the study of a certain empirical phenomenon, but rather to a pair composed by a phenomenon and a question about it. Such methods are used when it is impossible to identify a small number of independent variables from which values those of all other pertinent variables depend, so that their measurement is enough to describe the evolution of the phenomenon in a way relevant to answer the question at hand. We have argued that, when this happens, no appropriate understanding of the phenomenon is available, regardless of how one could conceive the notion of understanding. If, in spite of this, scientists can still offer answers to the questions posed, by utilizing methods that work with very large amount of (irreducible, or, at least, unreduced) data, then we say that these methods are "blind" and prefigure an instance of "agnostic science".

As to the need of understanding what makes the application of mathematics in agnostic science successful, we adopt, if possible, an even broader attitude. By advocating this need, we want merely to promote an informed research and reflection on what makes blind methods successful (when they are so), in spite of their blindness.

Let us now briefly describe the structure of the paper. We will begin by describing, in Section 2, what we consider to be the basic trend of big data, or the "Microarray Paradigm", as we called it in [9]. This name was chosen to reflect the fact that this trend became first manifest in biology and biomedical sciences, though it is now pervasive in all data analysis. It is characterized by the handling of huge amounts of data, whose specific, or at least fine-grained, provenance is often unknown and even practically unknowable, and whose modes of selection are often disconnected from any previous identification of a relevant structure in the phenomenon under observation. Far from being considered as a shortcoming, this

feature is intended as the most notable virtue of the paradigm, since it allows investigating the relevant phenomena, or better the data that have been gathered about them, without any previous hypothesis about the way they could be ordered, and the correlations and/or invariances that they would be supposed to confirm or refute.

There is, however, an important distinction to be done between the use of powerful and often uncontrollable algorithms on huge amounts of data and agnostic science properly said. This distinction will be investigated in Section 3 with the help of a negative example, the Page Rank algorithm used by Google to weight web pages. The basic point, here, is that the lack of local control of the algorithm in use is not the same as lack of understanding of the relevant phenomenon. The former is proper to any algorithm working on huge amounts of data, like Page Rank; the latter is instead, by definition, the characteristic feature of agnostic science.

In Section 4, we will go further in our analysis of agnostic science, by investigating the relations between optimization and "forcing". As we have argued in [11], optimization is at the very basis of what blind methods in data analysis pursue. In our view and terminology, this can be seen as a form of forcing: in absence of understanding of the relevant phenomena, and in order to solve a related problem, mathematics is forced over the available data; this is done, basically, by reiteratively appealing to interpolation, in order to identify a family of fitting functions within a functional space chosen for the operational advantages of its functions, and to optimization, in order to successively improve these functions' correspondence to the data. By better describing this modus operandi, in relation to deep learning techniques, we will also evaluate a very natural conjecture: that the effectiveness of optimization is the basic reason behind the success of agnostic science. We will, however, argue against this conjecture, once again noting that optimization is itself a form of forcing that does not ensure that the fitting functions and their extrema correspond to anything of significance in the evolution or state of the phenomenon.

Once this conjecture is rejected, we are still left with the question of the success of agnostic science. In our final Section 5, we shall outline a tentative answer by indicating a possible direction for further reflection.

2 The Microarray Paradigm

Let us focus on a DNA microarray (one can find several images and descriptions on the web). What is it? How does it work?

A DNA microarray is essentially a matrix of microscopic sites where several thousands of different short pieces of a single strand of DNA are attached. Messenger RNA (mRNA) molecules are extracted from some specific tissues of different patients, then amplified and marked with a fluorescent substance and finally dropped on each site of the microarray. This makes each site take a less or more intense florescence according to the amount of mRNA that binds with the strands of DNA previously placed in it. The intensity and distribution of the fluorescence give a way to evaluate the degree of complementarity of the DNA and the mRNA strands.

We know that the specific behavior of a cell largely depends on the activity, concentration, and state of proteins in it, and the the distribution of proteins is, in turn, influenced by the changes in levels of mRNA. This provides a correspondence between the information displayed by a DNA microarray and the behavior of a cell from the relevant tissue. This correspondence, however, is by no means exact or univocal, since the function of many proteins in the cell is not known, and several strands of DNA are complementary to the mRNA strands of all protein types. Nevertheless, thousands of strands of DNA are checked on a single microarray, so that one might expect this method to offer a fairly accurate description of the state of the cells. Such a description does not offer, however, any sort of understanding of what is happening in the relevant tissues, since the microarray supplies a particular value for a huge number of variables, whose relation to each other and to the state of the cell we ignore.

This lack of understanding does not forbid using microarray in the care (and, possibly, prevention) of many illnesses, since by measuring the activity of proteins one can hope of distinguishing patients and tissues affected or not by a certain pathology or reacting or not to a certain therapy, even without knowing why this is so.

This short description should be enough to justify why we take microarrays as a paradigmatic example of the way agnostic science works. What we call 'microarray paradigm' could be shortly described by the following slogan: *if enough and sufficiently diverse data are collected regarding a certain phenomenon, we can answer all relevant questions about it.*

But how much data is enough, in order to get a reliable conclusion? This depends both on the phenomenon and the specific question asked about it. But there are ways to apply data science also to small data systems if we accept strong limitations on the type of questions and we impose strong restrictions on the type of solutions (essentially regularizations constraints), providing fast optimization.

This already displays what we mean by forcing. But before coming to it, let us emphasize the general schema which is applied here: the data are processed through a blackbox, and the process itself is independent both of the specific nature of the data, and of any knowledge (or even hypothesis) on the role, mutual dependence and, more generally, relations of the variables these datas provide values of. The process is subject to to normalization constraints imposed by the data, rather than by the structure of the phenomenon (which is indeed unknown). This treatment produces an output which is taken as an answer to a specific question about this phenomenon.

This approach makes it impossible to generalize the results or even to deal with a change of scale: any question is answered by a different use of an algorithm, or a different algorithm. What is general is the structure of each algorithm, but not its use. One can say that the way in which a question is formulated depends on the nature of the algorithm which is used, and not the other way around.

To better see how agnostic science works, we will offer an overview of Supervised Machine Learning (we shall came back later to this description in more detail).

The starting point is a training set (X, Y), constituted by M pairs (X_i, Y_i), $(i = 1, 2, \ldots, M)$, where each X_i is typically an array $(X_{i,j})$, $(j = 1, 2, \ldots, N)$ of given values of N variables

pertaining to the same sample, for example a single patient in the case of a diagnostic test, and each Y_i is the relevant given output for this sample, for example either 0 or 1 for each patient according whether this patient is sick or healthy.

We can then use this set to find, by an appropriate algorithm, a suitable function F such that $F(X_i) = Y_i$ or $F(X_i) \approx Y_i$ $(i = 1, 2, \ldots, M)$. Here 'suitable' does not merely mean that F satisfies these conditions for all relevant i, but that it does it on most cases, and that it belongs to a space of functions selected because of its mathematical simplicity (for example the space of polynomial functions).

This function F is, then, tested on a testing set (X_i, Y_i), $(i = M+1, M+2, \ldots, M+M')$, and after having been suitably modified is finally used, by analytical continuation, as a base for forecasting.

This description indicates that Supervised Machine Learning consists, essentially, in an analytic continuation of a function found by interpolation, within an appropriate functional space. What is relevant, however, is that each of the numbers M, N, and M' may be huge, and we have no idea of how the values Y_i depend on the values X_i, or how the values in each array X_i are related to each other. In particular, we do not know whether the variables taking these values are reducible (that is, depend on each other in some way) or whether it is possible to apply any appropriate changes of scale or normalization on the variables. As anticipated above, this is just what we mean by speaking of lack of understanding: we have no other resource than using appropriate algorithms to identify a possible interpolating function F. We should add that the interpolation algorithm usually depends on parameters used to weight the available data. Still, because of the lack of understanding, there are no real criteria to guide the choice of the parameters, which are instead taken with arbitrary values, eventually corrected by successive reiterations of the algorithm, until some sort of stability is achieved.

This process seems to raise an obvious question. Indeed, using Ramsey's theory, Calude and Longo ([3]) have shown that if one enlarges sufficiently the training and the testing sets, it is possible to establish any possible correlation among data. Thus, by working on larger and larger amounts of data, one can establish arbitrary descriptions and arbitrary forecasts. In other words, large enough data seem to allow the establishment of any possible interpretation. If this is indeed the case, then the use of big data is doomed to failure.

There are, however, several comments to be made. To begin with, we note that Ramsey's theory proves the existence of lower bounds on the size of data in order to find non-arbitrary (if not structural) correlations. But this is still not enough to ensure that these bounds are small enough for becoming significant with respect to the present possibility of handling data. Even more importantly, we note that requiring, as in supervised machine learning, that every element of X match with an appropriate element of Y is essentially different from making a subset of the data display a certain correlation. In other words, if, in line with the tenants of Ramsey's theory, an algorithm showed that only for some subset of elements of X and Y it is possible to write $F(X_i) = Y_i$, this would have no useful application in practice for supervised learning, where the totality of the available data is required to be properly matched, by reducing the training and testing error rates as much as possible. Of course

there is always the risk that, even after learning from large training and testing sets, the fitting function F has no predictive power, but to identify such a problem, it is not necessary to invoke Ramsey theory; this is, indeed, the consequence of the obvious limitation of any interpolation procedure, and of any empirically validated rule.

Hence, as long as finding patterns within a data set X is tied to supervised learning, i.e. to the solution of a predetermined problem defined by the suitable matching of X and Y, it will not be subject to the risks of uncontrolled and spurious correlations. Moreover, we will see in Section 4 that, even when agnostic methods seem not to fall within the structural constraints of supervised learning, they can still be reinterpreted as such, for the specific problem of regularizing or compressing data.

This does not mean that we can avoid all risks of arbitrary descriptions and forecasts, risks that are indeed often present in agnostic science. Rather, agnostic science enters the game not in opposition to traditional, theoretically-tied methods, but as an other mode of exploration of phenomena, and it should be in no way (used to) discourage, or even inhibit the search for other sort of methods based on previous understanding. Any form of understanding of the relevant phenomena is certainly welcome. Still, our point here is that there is no intrinsic methodological weakness in blind methods that is not in a way or another already implicit in those other methodologies with a theoretical bent: at their core they all depend on some sort of inductive inference, namely on assuming that a predictive rule, or a functional interpolation of data, either justified by a structural account of phenomena, or by analytical continuation of functions got by interpolation on huge amount of (independent) data, will continue to hold true when confronted with new observations.

Supervised learning shows that we can succeed, despite the obvious (theoretical and/or practical) risks, in using data to find pattern useful to solve specific problems with the available resources (though not necessarily patterns universally associated with the relevant phenomena). This, and only this (namely not any alleged infallibility or omnipotence of them) makes (or, at least, should make) agnostic science both useful and welcome.

It is also because of these risks that it is important to understand why agnostic science works and what makes it successful. We should not be blind about why blind methods succeed! Lack of understanding of phenomena does not necessarily require or even allow lack of understanding of agnostic science itself. Rather, it urgently asks for such a latter understanding, in order to allow some sort of indirect (scientific, methodological, political and ethical) control possible, if not to allow any risk in the use of blind methods. This is just the aim of an informed philosophy of data analysis, which shows, then, not only its intellectual interest, but also, and overall, its practically utility, even necessity.

3 Agnostic Science and Lack of Control

Before continuing our search for such a (meta-)understanding, a proviso is essential. One might fall into conflating agnostic science with the mere use of uncontrolled algorithms on huge amount of data. Indeed, powerful algorithms can be applied on huge amounts of data whose size makes it (practically) impossible to exercise any sort of local control on the

way these same algorithms work, even when it is perfectly clear (even provably clear) that they converge, we understand what they result in, and even when the result depends on a structural understanding of the relevant phenomenon. This would not be an example of agnostic science. To see how this is possible, take the example of PageRank: the algorithm used by Google to weight webpages ([2], [12]).

Let A be a web page with n other pages T_i $(i = 1, 2, \ldots, n)$ pointing to it. We introduce a damping factor d_A $(0 \leq d \leq 1)$ that describes the probability that a random web-surfer landing on A will leave the page. If $d_A = 0$, no surfer will leave the page A; if $d_A = 1$, every surfer will abandon the page. One can chose d_A arbitrarily, or do it on the base of any possible *a priori* reason. This makes no difference in the end, or at least at the limit. The PageRank of A is given by this formula:

$$PR(A) = (1 - d_A) + d_A \left(\sum_{i=1}^{n} \frac{PR\left(T_i\right)}{C\left(T_i\right)} \right).$$

where $PR\left(T_i\right)$ and $C\left(T_i\right)$ are the PageRanks of T_i and the links going out of it.

This formula is very simple. But it is recursive: in order to compute $PR(A)$, one needs to compute the PageRank of all the pages pointing to A. In general, this makes impossible to directly compute it, since, if A points to some T_i, then $PR\left(T_i\right)$ depends on $PR(A)$, in turn. This does not make the computation of $PR(A)$ actually impossible, however. Since one can compute it by successive approximations: one begins by computing $PR(A)$ by choosing any arbitrary value for $PR\left(T_i\right)$; the value of $PR(A)$ so computed, is, then, used to provisionally compute $PR\left(T_i\right)$; next one comes back on $PR(A)$ to compute it anew on the base of these values of $PR\left(T_i\right)$; and so forth, for a sufficient number of times.

It is impossible to say *a priori* how many times the process is to be reiterated in order to reach a stable value for any page of the Web. The actual complexity and dimension of the Web makes moreover impossible to follow the algorithm's computation in any of its stages, and for all the relevant pages, even for a single page A, if the page is sufficiently connected within the Web. Again, since the Web is constantly changing, the PageRank of each page is not fixed and is to be computed again and again, so that the algorithm can never really stop to run. Thus it is obvious the impossibility of any local control on this process.

Still it can be demonstrated that the algorithm converges to the principal eigenvector of the normalized link matrix of the Web. This makes the limit PageRank of any page, namely the value of the PageRank of the given page in this vector, a measure of the centrality of this page in the Web. This does not mean, of course, that it also measure the actual importance of the page. Still, this is an entirely different story. What is relevant is that the algorithm have been designed to compute the principal eigenvector, under the assumption that the value obtained in this way is an index of the importance of the page.

Masterton, Olsson and Angere ([7]) have, by the way, recently proved that it is sufficient to suppose that a link pointing to each web page, in whatever portion of the Web, is somehow motivated by the importance of this page—however we define importance—and not by the fact that other links point or not to it, to make the PageRank of the pages in this portion of the Web actually converge to the importance of these pages when the size of this portion

increases. Let us tell it in another way: suppose that a measure of importance is ascribed to each page in this portion of the Web in whatever possible way, and that, for whatever such page, the probability that another page points to it increases with such a measure according to whatever monotonically increasing function, but is independent of the probability that any third page point to it; then the probability that the ranking of the pages in this portion of the Web assigned to them by PageRank coincides with the ranking due to their importance converges to 1 when the size of the Web goes to infinity.

This result confirms the essential point: the algorithm responds to a structural understanding of the Web, and to the (motivated) assumption that the importance of any page in it is proportional to its centrally in its normalized link matrix. Then, strictly speaking, there is nothing blind in this approach, and using it is in no way an instance of agnostic science, though the Web and the net of links in it is one of the most obvious example of Big Data we might imagine. But, then, what makes blind methods blind, and agnostic science agnostic?

Agnostic science appears when, for the purpose of solving specify problems, one uses methods to search invariants which—unlike PageRank—correspond to no previous understanding. This means we use methods and algorithms to find problem-dependent invariants in the hope that, once discovered, they will provide an apparent solution to the given problem.

If this is so, then agnostic science is, in fact, a family of mathematically sophisticated techniques to learn from experience by observation of invariants. Still, attention to invariants is, ultimaely, that which Plato (*Theaetetus*, 155d) called 'astonishment [θαυμάζειν]', and considered to be "the origin of philosophy [ἀρχὴ φιλοσοφίας]". What happens with agnostic science is that we know too much, or not still enough on too much data, in order to be astonished by our experience as guided by the conceptual schemas we have at hand. So we use blind methods to look for sources of astonishment deeply hidden within these data.

4 Forcing

The question we tackle is, then, that of understanding what makes an algorithm able to identify appropriate invariants within a specific problem.

The question has two facets. On one hand, it consists in wondering what makes these algorithms successful. On the other hand, it consists in wondering what makes them so appropriate (for a specific problem). The problem is that what appears to be a good answer to the first question seems to contrast, at least at first glance, with the possibility of providing a satisfactory answer to the second question.

Indeed, as to the first question, we would say that the algorithms perform successfully because they act by forcing, i.e. by choosing (interpolation) methods and selecting functional spaces for the fitting functions in agreement to a criterion of intrinsic (mathematical) effectiveness, rather than building them into the relevant phenomena. Is this answer really in contrast with the possibility of providing a satisfactory answer to the second question? In this section we shall try to make this answer more precise and to explore ways to make it com-

patible with the second question. Unfortunately we will see that none of these approaches is promising, and so in the next, final section, we will suggest a different direction.

A clear example of forcing is given by boosting algorithms: algorithms designed to improve weak classifiers, generally just slightly better than random ones, and to transform them, by iterations in strong classifiers. A second example is given by regularization algorithms. If the data are too complicated and/or rough, these algorithms are designed to render the date amenable to being treated by other algorithms, for example by reducing their dimension. Using these algorithms reveals a double application of forcing: forcing on the original data to smooth them; and then forcing on the smooth data to treat them with a second set of algorithms. After regularization, if this is needed, one uses often continuity and smoothness conditions to force an essentially unjustified differential equation in the search for a fitting function ([15], chapter 19). Again, it often happens that methods originated in a certain domain (where they are proved to be highly effective) are exported to other, completely different and unrelated domains. A case in point are neural networks used, for example, in studying climate change.

In all these cases, the mathematical ~~hard~~ core of the methods is provided by some sort of optimization techniques. Our claim is, then, that optimization can be see as a form of forcing. In a sense, this is even suggested by the historical origins of optimization methods ([13]; [14]). When Maupertuis firstly introduced the idea of least action, he claimed to have found the real quantity that God had aimed to minimize when creating the universe. Euler, who was at the time a member of the Berlin Academy of Sciences, whose President was Maupertuis himself, could not openly criticize his President, but clearly adopted a different attitude, by maintaining that action was nothing but what was expressed by the equations governing the system under consideration, provided they were shaped in an appropriate form. In other terms, he suggested one should force the minimization (or maximization) of

$$\int F(x)dx$$

on any physical system in order to find the function F characteristic of it. *Mutatis mutandis*, this is the basic idea that we associate today with the Lagrangian of a system. Since then optimization became the preeminent methodology in solving empirical problems. One could say that the idea of a Lagrangian has been generalized to the notion of fitting function, whose optimization characterizes the dynamics of a given system.

Though this process can be already seen as a form of forcing, acting within a quite classical setting, one should note that in this case the only thing that is forced on the problem is the form of the relevant condition, its being shaped as the request that a certain appropriate integral reach a maximum or minimum. Since, the function to optimize is here so chosen as to express at least a preliminary understanding of the system itself. Things change radically when the fitting function is selected within a convenient functional space through an interpolation process designed to have the function fit the existing data. In this case, both the space and the nature of the fitting (that which makes it appropriate) are forced on the system. Moreover, often these conditions are still not enough to select a unique fitting function or to find or ensure the existence of an absolute minimum, so that a

new functional choice may be required, and thus forced again on the data.

There are many reasons why such an optimization process can be considered effective. One is that it matches with what we have called before the microarray principle. Enough data and a sufficiently flexible set of algorithms will solve, in principle, any scientific problem. More concretely, optimization has shown to be both simple and relatively reliable: not necessarily to find the actual solution of a problem, but rather to obtain, whiteout exceeding time and resources constraints, outcomes that can be taken as useful solutions to the problem. These outcomes can be also tested in simple cases and shown compatible with already known solutions found with other methods based on a structural understanding of the relevant phenomenon; in addition the results from the optimization process may be suitable for practical purposes (as in the case of self-driving cars, where the aim is not that of mimicking human reactions, but rather that of having a car driven with appropriate care). Finally, we can conceive optimization as a motivation for finding algorithms without being constrained by the searching of the best solution.

But optimization as forcing also raises some important issues, beyond the obvious one which is typical of blind methods, namely the absence of an *a priori* justification.

One is that optimization generally requires fixing a large number of parameters, sometime millions of them, which not only make control hopeless, but also makes it difficult to understand the way the algorithm works, and often results in lack of robustness, since different initial choices of the parameters can lead to completely different solutions. Another is manifest in point by point optimization, for example by the gradient descent method ([4], section 4.3), which does not guarantee that we will reach the desired minimum, or even a significant relative minimum. Since virtually all significant Supervised Machine Learning methods can be shown to be equivalent to point by point optimization [11], we will describe and discuss the gradient descent method.

If $F(X)$ is a real-valued multi-variable function, its gradient ∇F is the vector that gives the slope of its tangent oriented towards the direction in which it increases most. The gradient descent method exploits this fact to obtain a sequence of values of F which converges to a minimum. Indeed, if

$$x_{n+1} = x_n - K_n \nabla F\left(x_n\right) \quad (x = 0, 1, \ldots)$$

for K_n small enough, then

$$F\left(x_0\right) \geq F\left(x_1\right) \geq F\left(x_2\right), \ldots$$

and one can hope that this sequence of values converges towards the desired minimum. This is only a hope, however, since nothing in the method can warrant that the minimum it detects be significant, and, even less, appropriate for an optimization apt to provide a solution for the relevant problem.

Let us now further illustrate the idea of optimization as forcing, by considering the paradigmatic example of Deep Learning Neural Networks (we follow here ([5], Section 11.3) and [4]).

The basic idea is the same anticipated above. One starts with a Training Set (X, Y),

where X is, as a matrix (an array of arrays) of variables,

$$X = (X_1, \ldots, X_M) \qquad X_i = (X_{i,1}, \ldots X_{i,N})$$

while Y can be either an array of variables,

$$Y = (Y_1, \ldots, Y_M)$$

which is the case when attempting some data classification, or an array of array of variable, in more general cases. The goal is to find a function $F : X \to Y$ in a certain functional space, whose optimization will give us the desired result. The functional space is chosen so that any function in it depends on parameters, which weight the data in X. The space will include all the functions obtained by composition, according to a certain algorithm, of some basic functions whose general form is fixed in advance (and varies slightly from a particular version of the method to another). As the algorithm moves from a function to another, the choice of the parameters also changes, and this is what the term 'learning' means.

The process starts with K linear functions for any array X_i in X:

$$Q_i^{[k]}(X_i) = A_0^{[k]} + \sum_{j=1}^{N} A_j^{[k]} X_{i,j} \qquad (k = 1, \ldots, K),$$

where $A_j^{[k]}$ are $K(N+1)$ parameters chosen with base on same *a priori* criterion, or even randomly, and K is a positive integer appropriately chosen in relation to the particular application of the algorithm. Then, one gets K new arrays of variables

$$H_i^{[k]} = G\left(Q_i^{[k]}(X_i)\right) \qquad (k = 1, \ldots, K),$$

were G is an appropriate non-linear function, changing from a particular version of the method to another: we will soon say more about this function. Next, one gets T linear combinations of the variables $H_i^{[k]}$

$$Z_i^{[t]}(H_i) = B_0^{[t]} + \sum_{j=1}^{K} B_j^{[t]} H_i^{[j]} \qquad (t = 1, \ldots, T),$$

where $B_j^{[t]}$ are $T(K+1)$ parameters chosen with base on same *a priori* criterion, or even randomly, and T is a positive integer appropriately chosen in accordance with the particular application of the algorithm.

If the neural network has only one layer, that is, it is not deep, we set $T = 1$ and the process closes by imposing

$$Z_i^{[1]} \approx Y_i$$

and modifying the values of the parameters accordingly, in a way to be explained soon. As neither the algorithm nor the parameter depends on i, this is the same as choosing $Z^{[1]}$ (the same for any i) as our final function $F : X \to Y$. If the network is deep, the process is repeated several times, starting from the M arrays Z_i each of length T, to produce several

layers, by choosing different parameters A and B at each step, with the obvious limitation that the dimension of the output of the last layer has to match the dimension of the elements of Y.

The algorithm is designed to allow learning also in absence of Y by using X itself, possibly appropriately regularized, in place of Y (auto-encoding). When an independent Y is used, the learning in said to be 'supervised' and corresponds to the setting described in § 3. In absence of it, it is equivalent to unsupervised learning ([4], chapter 14), where the objective is to find significant patterns and correlations within the set X itself. This is an important shift of perspective, because it allows to constrain the exploration of patterns within data X, for the sole purpose of regularization of the data themselves. Whichever correlations and patterns are found, they will be instrumental to this specific problem, rather than on the ambiguous task of finding causal relationships within X.

Two things remains to be explained.

The first concerns the non-linear function G, called 'activation function' (because of the origin of the algorithm as a model for neural dynamic). It can take different forms. Two classical examples are the sigmoid function

$$G(u) = \frac{1}{1 + e^{-u}}$$

and the ReLU (Rectified Linear Unit) function

$$G(u) = max(0, u),$$

This latter function is composed of two linear branches and therefore is, mathematically speaking, much simpler than the sigmoid function. While the ReLU is not linear, it has a steep uniform slope, on a wide portion of its domain, and this seems the key of its significantly better performance as activation function for deep networks. The use of an appropriate activation function allows the method to approximate any nice (continuous on closed and bounded subsets of the n-dimensional real space $\mathbb{R}^n$) function. This is the Universal Approximation Theorem for Neural Networks ([6])

The second thing to be explained concerns the computation of the parameters according to the condition

$$Z_i \approx Y_i.$$

This is typically made through the gradient descent method (with the gradient computed by an appropriate fast algorithm adapted to neural networks known as backpropagation: [5], Section 11.4), with the request to minimize

$$\sum_{i=1}^{M} \left[Y_i - Z_i\right]^2.$$

To this purpose, one can go as far as considering hundreds of layers, though it is not generally true that increasing the number of layers always improves the minimum, nor it makes Z_i closer and closer to Y_i. All what can be said, in general, is that once one fixes the number

of layers, the algorithm of gradient descent used to establish the relevant parameters is equivalent to a regularization of the final function, if we stop the iterative application of this algorithm when the error does not significantly decreases any more ([4], Section 7.8). Still, in many cases, taking more layers makes a dramatic improvement possible: up to a tenfold reduction of errors. For example, it has been showed that in a database of images of handwritten digits classification errors go from a rate of 1.6% for a 2 layers network ([5], section 11.7) to a rate of 0.23% with a network of about 10 layers ([16]).

This short explanation of the way in which Deep Learning Neural Networks work should be enough to clarify why we have taken them as an example of optimization by forcing. But it should also be enough to make clear the second aspect of the question mentioned in the beginning of the present section: how can methods like this be appropriate for solving specific problems, when the methods themselves do not reflect in any way the particular features of the problems?

A simple way to answer is by negating the premise of the question: one can argue that, in fact, these method are in no way appropriate; that their success is nothing but appearance and that, rather, the faith in their success is dangerous, since it provides an incentive to the practice of accepting forecasts and solutions that are indeed inappropriate or erroneous.

The problem with this answer is that it argues that blind methods do not succeed since they cannot succeed because they do not conform with the pattern of classical science. But a non-ideological look at the results obtained in this way should be enough to convince ourselves that this cannot be the good answer. We can of course emphasize, for example, that basing cancer therapy only on microarrays, and using microarrays only as guides for therapies is as inappropriate as dangerous, since, in a domain like that, looking for causes is as crucial as necessary. But we cannot hide the fact that microarrays can be used also as an evidential bases in a search for causes. And also that, in many cases—like in handwriting recognition—the search of causes is much less crucial, and that in many situations the success of blind methods is manifest and confirmed by applications.

So we need a less ill-advised answer, which, far from negating the question, takes it seriously and challenges the assumption that classical science is the only appropriate pattern for good science.

Such an answer cannot depend, of course, on the assumption that blind methods succeed since they perform appropriate optimization. And this not simply because this merely displaces the problem, but mostly because, in these methods, optimization is only such by name and not by nature. Indeed, when forced on a problem, optimization can be identified, in a general sense, with the mathematical form of classical optimization, but certainly not because it actually reaches an optimum in the sense of finding the best function or the absolute minimum of the right function.

A more promising answer might be that blind methods succeed for the same reason as classical induction does: blinds methods are indeed interpolation methods on income/outcome pairs, followed by analytical continuation, which is the basic way in which induction works. Of course, one could argue that induction itself is not logically sound. But could one really reject it as an appropriate method in science because of this? Is there another way to be em-

piricist, other than trusting induction? And can one really defend classical science without accepting some form of empiricism, as refined as it might be?

One could offer two important objections to this response.

The first is that our argument applies only to supervised methods, that is methods based on the consideration of a training set, and, possibly, also a testing one. It does not apply, or, at least, not immediately, to unsupervised ones, where no sort of induction is present. One could supersede this objection, however, by noticing that it is possible to reduce unsupervised methods to supervised ones through the auto-encoding regularization processes described above.

A second objection is much more relevant. The objection consists in recognizing that when forcing is at work, interpolation is restricted to a space of functions which is in no way selected by a consideration of the specific nature of the relevant phenomenon and, in view of that, it is submitted to a regularization of the data that often takes those same data quite far from reflecting the phenomenon itself. Because of this objection, and despite the fact that the answer we offered cannot be completely dismissed, it is necessary to complement it with a more specific and stronger response.

5 Conclusions

Should we, then, abandon any hope to provide a satisfactory answer to our question, capable to reconcile its two aspects, that is, able to explain at once the good performance of blind methods and their appropriateness for the solution of the problems they are applied too? Or should we abandon the answer we have given with respect to the first aspect, namely the claim that the relevant methods operate by forcing? In other words, should be accept that forcing (or, more generally, blindness) is incompatible with appropriateness? We think not. Not only, we are not ready to renounce, but, though we are far from having ready a totally convincing, exhaustive and definitive answer, we would like to suggest a new perspective, which we believe will be fruitful.

The basic idea is to stop looking at the appropriateness question as a question concerned with some sort of topical correspondence between phenomena (in particular the relevant problems about them) and methods. The very use of forcing makes illusory the possibility to identify such a correspondence.

We should instead look at the question from a more abstract, general or, as it were, structural perspective. Why not to imagine that what makes blinds methods fits in some way or another with the phenomena and problems they are successfully applied to is a sort of homology of structure between the former and the latter?

If we look at blind methods from a more structural point of view, we can find a simple, general feature that is shared by all of them. It relates to what we have elsewhere ([11]) called 'Brandt Principle'. We called it this way, since we did find it firstly expounded, though implicitly and for the restricted class of multiscale algorithms, in a paper by Achi Brandt ([1]) and can be stated as follows:

> An algorithm that approaches a steady state in its output has found a solution to a problem, or needs to be replaced.

As trivial as it appears at first glance, this is, in our view, the fundamental principle on which agnostic science is founded. First of all, the principle is implicit in forcing, since an integral idea in forcing is that if an algorithm does not work, another one is to be chosen. But, more specifically, the key to the power of this principle is that the steady state output of each algorithm, when it is reached, is chosen as input of the next algorithm, if a suitable solution to the initial problem has not yet been found. Notably, deep learning architecture matches with it, since the reiteration of the gradient descent algorithm is generally arrested when the improvement of parameters reaches a steady state and, then, either the function that has been obtained is accepted and used for forecasting or problem-solving, or the algorithm is replaced by a new one, or at least re-applied starting from a new assignation of values to the initial parameters. Again, local optimization methods satisfy this principle. And since most, if not virtually all, algorithms in agnostic data science can be rewritten as local optimization methods, we can say that virtually all algorithms in agnostic data science do. Moreover, in [11] we argued that thinking about algorithms in terms of Brandt's principle often sheds light on those characteristics of a specific method that are essential to its success.

For example, the success of deep learning algorithms, as we have seen in the previous section, relies in a fundamental way on two advances: (1) the use of the ReLU activation function that, thanks to its constant slope for nonzero arguments, allows the fast exploration of the parameter space with gradient descent; and (2) a well defined regularization obtained by stopping the gradient descent algorithm when error rates do not improve significantly anymore. Both these advances took a significant time to be identified as fundamental to the success of deep learning algorithms, perhaps exactly for their deceiving simplicity, and yet both of them are naturally derived from Brandt's principle.

But, as fundamental as it might be in agnostic science, why would such a principle provide an answer to our question on the power of agnostic science? The answer is that the general structure revealed by Brandt's Principle is not necessarily proper to algorithmic procedures of methods. It can also be shared, at a sufficient level of abstraction, by the phenomena they apply to.

Indeed, in [11] we showed how Brandt's Principle is equivalent in spirit to an organizing principle for developmental biology, first proposed by Alessandro Minelli in [8], what he called the principle of Developmental Inertia. Such principle states that:

> Biological developmental processes typically go ahead by successive deviations from local self-perpetuation of cell-level dynamics.

We refer to [11] for a full justification and explication of such principle, but we note here that, just like Brandt's Principle for algorithmic processes, the principle of Developmental Inertia has a surprisingly powerful ability to conceptualize developmental biological processes.

Its homology with Brandt's principle is evident by associating steady-states of algorithms with self-perpetuating (and repeating) cell dynamics, and by associating switching

of algorithms (as advocated by Brandt's Principle) with deviations of self-perpetuating cell dynamics.

The only structural difference between these principles is that Brandt's Principle starts from non-steady states, and prescribes what to do when approaching steady states, while the principle of developmental inertia starts from local self-perpetuating dynamics, and establishes the necessity of deviating from them for a complex biological process to develop. But this is only a difference of emphasis, since both steady states and deviations are postulated as essential by both principles.

In light of this parallel between Brandt's and developmental inertia principles, the question becomes the following: is the success of methods in agnostic science due to the homology of their general structure (as revealed by Brandt's Principle) with the structure of the phenomena (and problems) that agnostic science deals with?

For the time being, we prefer to leave the question open. But we are confident that future research and reflection might provide an answer illuminating enough for helping in answering the other, more fundamental question that we have tried here to make clear and illustrate.

References

[1] A. Brandt, A., Multiscale Scientific Computation: Review 2001, in T. J. Barth, T. F. Chan, R. Haimes (eds.) *Multiscale and Multiresolution Methods: Theory and Applications*, Springer Verlag, Berlin-Heidelberg, 2002, p. 3-95.

[2] S. Brin & L. Page, The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 1998, **30**, pp. 107-117.

[3] C. Calude, G. Longo, The Deluge of Spurious Correlations in Big Data, *Foundations of Science*, 2017, **22**, pp. 595–612.

[4] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*,The MIT Press, 2016.

[5] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics* Springer; 2nd edition (2016).

[6] K. Hornik, Approximation capabilities of multilayer feedforward networks, *Neural Networks*, 1991, **4**, issue 2, pp. 251–257.

[7] G. Masterton, E. J. Olsson & S. Angere, Linking as voting: how the Condorcet jury theorem in political science is relevant to webometrics. *Scientometrics*, 2016 **106**, 3, pp. 945-966.

[8] Minelli, Alessandro. 2011. A principle of Developmental Inertia. In *Epigenetics: Linking Genotype and Phenotype in Development and Evolution*, eds B. Hallgrimsson and B. K. Hall. Berkeley, CA: University of California Press.

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

[10] D. Napoletani, M. Panza, and D.C. Struppa, Is big data enough? A reflection on the changing role of mathematics in applications. *Notices of the American Mathematical Society* 61, 5, pp. 485–490, 2014.

[11] D. Napoletani, M. Panza, and D.C. Struppa, Forcing Optimality and Brandt's Principle, in J. Lenhard and M. Carrier (ed.), *Mathematics as a Tool*, Boston Studies in the Philosophy and History of Science 327, Springer, 2017.

[12] L. Page, S. Brin, R. Motwani, & T. Winograd, The PageRank citation ranking: bringing order in the Web. Manuscrit to be found at http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf.

[13] M. Panza, De la nature épargnante aux forces généreuses. Le principe de moindre action entre mathématiques et métaphysique : Maupertuis et Euler (1740-1751), *Revue d'Histoire des sciences*, 1995, **48**, pp. 435-520.

[14] M. Panza, The Origins of Analytical Mechanics in 18th century, in H. N. Jahnke (ed.) *A History of Analysis*, American Mathematical Society and London Mathematical Society, s.l., 2003, pp. 137-153.

[15] J. Ramsay, B. W. Silverman, *Functional Data Analysis*,Springer; 2nd edition, 2005.

[16] J. Schmidhuber, Multi-column deep neural networks for image classification. CVPR '12 Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Pages 3642-3649, IEEE Computer Society, 2012.