Agnostic Science. Towards a Philosophy of Data Analysis

Marco Panza, Domenico Napoletani, Daniele Struppa

Agnostic Science.
Towards a Philosophy of Data Analysis

D. Napoletani[1], M. Panza [2] and D.C. Struppa [3]

[1] Corresponding author: Department of Mathematical Sciences
George Mason University, Fairfax, VA 20030
Phone:+1-703-993-4269
Fax: +1-703-993-1491
email: dnapolet@gmu.edu

[2] IHPST (CNRS, Univ. Paris 1 and ENS Paris)
email: Marco.Panza@univ-paris1.fr

[3] Department of Mathematics and Computer Science
Chapman University, Orange, CA 92866
email: struppa@chapman.edu

April 20, 2010

Abstract

In this paper we will offer a few examples to illustrate the orientation of contemporary research in data analysis and we will investigate the corresponding role of mathematics. We argue that the *modus operandi* of data analysis is implicitly based on the belief that if we have collected enough and sufficiently diverse data, we will be able to answer most relevant questions concerning the phenomenon itself. This is a methodological paradigm strongly related, but not limited to, biology, and we label it *the microarray paradigm*. In this new framework, mathematics provides powerful techniques and general ideas which generate new computational tools. But it is missing any explicit isomorphism between a mathematical structure and the phenomenon under consideration. This methodology used in data analysis suggests the possibility of forecasting and analyzing without a structured and general understanding. This is the perspective we propose to call *agnostic science*, and we argue that, rather than diminishing or flattening the role of mathematics in science, the lack of isomorphisms with phenomena liberates mathematics, paradoxically making more likely the practical use of some of its most sophisticated ideas.

Keywords: Methods of Computational Science, Philosophy of Data Analysis, Philosophy of Science.

Vitae

- Domenico Napoletani is an assistant professor of mathematics at George Mason University in Virginia. His research focus is in the development and methodological understanding of signal processing and data analysis algorithms for highly structured large data sets with an emphasis on network reconstruction and control, classification, denoising.

- Marco Panza is research director at the CNRS (Paris, UMR 8590: IHPST). He got his PhD in philosophy of sciences in 1990 at the university of Genova (Italy) and his habilitation at the EHESS of Paris. Before being appointed at the CNRS, he was teaching in Geneva, Nantes, Mexico City, and Barcelona. His research focus is in the history of early-modern mathematics and in philosophy of mathematics.

- Daniele C. Struppa is a professor of mathematics at Chapman University in California. Dr. Struppa received his PhD from the University of Maryland in 1981, and most of his work has been in the area of Fourier analysis and its applications to harmonic analysis, and systems of differential equations. More recent interests include issues of signal processing and pattern recognition.

1 Introduction: The Role of Mathematics in Data Analysis

Data: what is given. It is difficult to find a more pervading word in today's scientific practice. In every field there is a surge of data collection, remarkable not only for its size, unthinkable until recently, but especially for its *modus operandi*: streams of values of variables are collected from a given phenomenon, without the pretension of understanding how they can contribute to the explanation, or simply to a suitable general description of the phenomenon itself.

This *modus operandi* is implicitly based on the following, almost paradoxical belief: if we have collected enough and sufficiently diverse data, we will be able to answer any relevant question concerning the phenomenon itself. A striking and important example of such a trend can be observed in biosciences, where the effectiveness of drugs or the detection of diseases are approached, in practice, by studying clusters, similarities and other structured characteristics of huge arrays of chemical compounds (microarrays) derived by gene, protein or even tissue samples. Microarrays may be superficially seen just as one application of quantitative methods among many, but we believe instead that they are a paradigmatic example, and we shall term *microarray paradigm* the modus operandi that we highlighted above and which we can summarize as follows: *if we collect enough and sufficiently diverse data regarding a phenomenon, we can answer most relevant questions concerning the phenomenon itself.* Our point in choosing microarrays as emblematic is twofold: first of all, the microarray paradigm is not limited to biology, as we will explicitly show in Section 4. Moreover, by choosing microarrays as paradigmatic, we stress the obvious fact that biology is becoming one of the main engines of quantitative scientific developments, and of applied mathematics as well. The purpose of our paper is to clarify this principle and to discuss the way in which mathematics is used within the paradigm of science which goes with it.

* * *

In this paper we will offer a few important examples to illustrate the orientation of contemporary research in data analysis and we will investigate the corresponding role of mathematics. The methods we describe in Section 4, neural networks, boosting, automatic control, are generally considered a

form of statistical learning (or machine learning) [Hastie et al., 2001], to signify the automated, data-driven nature of these methods, and their ability to learn structures from the data. Some of them, like neural networks, have a long history and are very well established techniques, while others, like boosting, are very recent new developments that have not yet been explored in their philosophical implications. Our purpose is to show how all these methods are characterized by a weakness of purpose, an inability to provide general and appropriate models for the problems they are supposed to solve. Above all we ask ourselves whether these methods can provide the basis of a fruitful general methodology of data analysis and whether they present novel philosophical questions, or methodological possibilities, distinct from those generated by a more traditional way of doing science.

The examples we put forward show how the role played by mathematics in the solution of empirical problems is changing drastically. This change makes it possible for mathematics, even in its very sophisticated forms, to play a significant role in domains that are relatively new to a quantitative analysis, such as biomedical sciences and meteorology.

To better explain our argument, we note that, at the deepest and most general level, mathematics is used to find symmetries and invariants and therefore to give a structure to a particular phenomenon. The mathematician will then deduce more properties within the mathematical theory which describes these structures, starting from suitable premises and principles expressed in the language of this same theory. This traditional approach produces a sort of isomorphism between some aspects of the phenomenon and a mathematical structure; this isomorphism is not only responsible for accurate, appropriate and correct forecasts, but, more importantly, it is taken as a manifestation of the hidden nature of the phenomenon.

Current scientific practice in data analysis works differently. In this new framework, mathematics provides powerful ideas and techniques which then generate broad classes of generic computational tools. These tools will be applied to large data sets to find the solution of a given particular problem. While in some case the computational tools may be useful as an intermediate step between theories and phenomena, as stressed in [Humphreys, 2004] and [Humphreys, 1995], more often the way they are applied to the data is leading to a scenario where it is missing any explicit isomorphism between a mathematical structure and the phenomenon under consideration. The computational tools of data analysis therefore do not truly model the phenomenon.

According to [Donoho, 2000], this current trend in data analysis is temporary and depends on a heuristic exigency in the face of new difficult problems. The present approach will eventually be replaced by another, more traditional approach, which relies on new, yet undiscovered, theories. We do not share this view. We believe, instead, that this trend is revealing a new methodological paradigm, which characterizes data analysis to a large degree. Far from being the result of a temporary inability of creating isomorphisms between phenomena and theories, this paradigm depends on a new conception which is gradually imposing itself and which deserves to be understood and appreciated as such. Instead of attempting to understand and model a phenomenon, this paradigm suggests that a scientist needs to approach a phenomenon with a limited set of assumptions, and needs to look for specific techniques capable to solve some of the problems it presents, without attempting any sort of structural understanding of the phenomenon itself. From this point of view, each phenomenon requires its own description, that will vary according to the available type of data, that cannot and should not be generalized to other phenomena that may be apparently similar. The tools that we apply may be similar for broad classes of phenomena, but actual descriptions will be very sensitive to the individual characteristics and fluctuations of the data from each phenomenon.

The methods used in data analysis are suggesting the possibility of forecasting and analyzing without understanding (or at least without a structured and general understanding). More specifically, we will argue in Section 2 that understanding occurs when we know how to relate several descriptions of a phenomenon. Instead, these connections are disregarded in many data analysis methodologies, and this is one of the key features of modern data analysis approaches to scientific problems. The microarray paradigm is at work exactly when a large number of measured variables concerning a phenomenon are algorithmically organized to achieve narrow, specific answers to problems, and the connections among different levels of descriptions are not explored. This is the perspective we propose to call *agnostic science* and at its heart is the methodological principle that we called microarray paradigm. The notion of such a paradigm is not intended to fix a specific procedure for questioning data, but rather an epistemic modality of our relation to them and a modality of interaction between them and mathematics. In this sense, even when we will talk of established statistical learning techniques such as classification methods (cf. Section 4.1) and time series analysis (cf. Section 4.2), the focus is not on the technique itself, but on the way the technique is

applied in the context of the microarray paradigm.

Rather than reducing the impact of mathematics in science, this approach liberates mathematics from the constraints of strict isomorphisms; and it allows a plethora of theoretical tools to be used in very practical scenarios as we will explore in Section 5.

2 Scale and Understanding

For the purpose of this article, data analysis refers to the treatment of a class of scientific problems considered as pairs composed by a phenomenon and a specific problem about it. Concrete and emblematic examples of this situation include the problem of forecasting the evolution of an hurricane during a certain lapse of time, or the intriguing problem of determining distinct sound sources from microphone recordings of their mixtures, what is often called in the literature the *cocktail party* problem (cf. [Cichocki and Amari, 2002]).

Some of these phenomena are such that their structure is partially known, for example, the microscopic physics of a hurricane can be described by fluid-dynamics partial differential equations, and yet the specific questions which are asked about these phenomena resist any easy solution that relies on such structure. In the case of hurricanes, the global behavior of the hurricane cannot be easily deduced by the microscopic description, since, in practice, any useful description of the evolution of an hurricane would involve *ad-hoc* tuning of a great number of model parameters that are not implicit in the microscopic physics. In other cases, for example, the reconstruction of the independent sources of a mixed sound, we lack any clear and/or precise understanding of the phenomena themselves, since we do not know the sources that generated the mixtures.

In both these general settings, the solution of the (empirical) problem about the phenomenon under consideration depends, predictably, on the solution of a mathematical problem that is connected to it. A deeper look at these connections will indicate what frustrates our understanding of phenomena, and it will suggest that the very notion of understanding is intimately related to the multiscale nature of any actual phenomenon.

Consider again the study of hurricanes, here the phenomenon under consideration is relatively well known insofar as its knowledge depends on a representation by a system of differential equations derived from physics.

The solution of this system, however, goes beyond the limits of our actual mathematical abilities, and we are actually unable to effectively determine the evolution of the hurricane from first principles. A typical strategy in such cases is to look for a numerical simulation of the given system of differential equations. This means that one looks for the determination of enough meteorological values to allow a sufficiently precise prediction of the behavior of the hurricane. This process can be reduced to the problem of solving a system of linear equations $Ax = b$, where A is a square matrix whose size is very high, while b and x are two vectors, known and unknown, respectively. The size of A depends on the number of values we want to compute: the more fine-grained is this simulation, the larger the number of variables that the discretization of the system will generate in the corresponding system of linear equations (cf. [Trottenberg et al., 2000]), and, again, our computational abilities will be overwhelmed. In these cases, we need to resort to heuristic adjustments that are not easy to justify from the microscopic physics of the hurricane, as we pointed out earlier.

Let us now consider the cocktail party problem, i.e. the reconstruction of the independent sound sources hidden in the recording of a mixed sound. The data we consider in this case are the recordings of N long time series $B = [b_1, ..., b_N]$ from a collection of microphones, and we assume that there are M underlying sound sources $X = [x_1, ..., x_M]$, linearly mixed so as to produce the recorded time series. Hence the problem can be modelled, once again, though a system of linear equations $AX = B$ where B now is a known matrix and we have to find simultaneously the unknown matrices A and X. This problem can be seen as the attempt of solving simultaneously several linear equations, though we lack the relevant information to solve them uniquely, and it can be solved only by making weak, but subtle assumptions on A and X (cf. [Lee, 1998]). These methods are so successful that they are called blind methods [Cichocki and Amari, 2002], since they seem to be able to solve problems without the necessary information. Their importance is growing especially in signal processing and image processing, where it is difficult to model the processes that generated the recordings and the images, and where blurring and noise are a severe problem.

We can see that both for the mathematical problem derived from hurricane simulation, and for the cocktail party problem, the size of the matrix A depends on the precision we want to reach. In the second example, the size of A determines the number of distinct sources that we want to identify, in the first case, it determines the accuracy of the simulation. The technical term

for speaking of such accuracy is *scale*. The notion of scale is usually involved both when one is referring to the size of the discretization of the phenomenon (as in the case of the hurricanes), and when one is referring to the number of quantities or parameters involved in a computational description of such a phenomenon (as in the case of mixed sounds). In both cases, one speaks of fine and coarse scale: in the first case to indicate, respectively small and the large sizes; in the second case, to indicate large or small numbers of parameters.

* * *

We have until now purposely used the term 'understanding' without a specific definition. Below we specify to some extent how we suggest to conceive understanding in the cases we are interested in. However, our aim is not to provide any detailed account of understanding for natural and social phenomena. We mostly appeal to the notion of understanding negatively, as lack of it, and all that we need is therefore a term of comparison for our account of forecasting and analyzing without understanding. In principle, such term of comparison could be provided by some of the accounts of scientific understanding that can be found in literature, often in connection with explanation (cf., among others, [Scriven, 1962], [Toulmin, 1963], [Friedman, 1974], [Achinstein, 1983], [Salmon and Kitcher, 1989] [Weber, 1996], [Trout, 2002], [Regt and Dieks, 2005])[1]. However, most of these accounts are too specific for our purpose, since they make understanding dependent on some specific logic or epistemic characters of our theories, models, or, more generally, our modes of acquaintance with the relevant phenomena, like deductive power, entailment, unification, causality, reduction, minimality, representation of regularities, necessity, abstraction, familiarity, generality, simplicity, etc. To use these accounts of understanding as terms of comparison in our work would wrongly suggest that the contrast we would like to emphasize depends on their specific characters, whereas the lack of understanding we talk about is much more general.

For example, De Regt and Dieks' recent account ([Regt and Dieks, 2005]) insists on the opposition between scientific understanding and algorithmic

[1] A few of these accounts specifically point to the way mathematics enters understanding of natural and social phenomena as in works such as [Batterman, 2002] and [Morrison, 2006]; [Batterman, to appear] includes a critical discussion of more recent accounts of the explanatory role of mathematics for empirical phenomena and it suggests an alternative account as well.

methods, or, more generally, calculations endowed by same appropriate mathematical apparatus. They propose the two following criteria (*ibid.*, pp. 150-151):

- A phenomenon P can be understood if a theory T of P exists that is intelligible (and meets the usual logical, methodological and empirical requirements).
- A scientific theory T is intelligible for scientists (in a context C) if they can recognize qualitatively characteristic consequences of T without performing exact calculation.

The relevant theories that De Regt and Dieks are concerned with are clearly supposed to integrate a mathematical apparatus [2], and their account suggests that the algorithmic and/or calculative procedures endowed by such an apparatus do not enter science to provide understanding, but are rather guided by a previous understanding that mathematical theories help to provide on top of algorithmic aspects. This anti-algorithmic bias is the specific reason this account of understanding is not appropriate to provide the term of comparison we need in our work, since it suggests in turn that the opposition between understanding and forecasting we would like to emphasize [3] depends on the fact that understanding is conceived in an anti-algorithmic way.

[2]This is confirmed by footnote 7 (*ibid.*, p. 167) where they write: "If one wants to apply our analysis to non-mathematical, qualitative theories, we suggest to replace 'exact calculation' by 'complete logical argumentation' ". Moreover, by dealing with the example of Boltzmann "qualitative analysis" of the behavior of gases under the effect of temperature increasing (*ibid.*, pp. 152-153), they remark that the purpose of this analysis "is to give us *understanding* of the phenomena, *before* we embark on detailed calculations [...][which] are subsequently motivated, and given direction, through the understanding we already possesses". And they add: "Exact mathematical techniques [...] are obviously essential in modern science. What we emphasize in the importance of understanding as an *additional* epistemic aim of science".

[3]In his well-known 1974 paper on explanation and understanding ([Friedman, 1974], p. 8), M. Friedman has remarked that "to have grounds for rationality expecting a phenomenon is not the same thing as to understand it", and mentioned, as an example of it, indicator-laws that make us able "to predict some phenomenon on the basis of [...] initial conditions" without requiring "understanding of why the phenomenon occurred". Unfortunately, he has not developed this point neither in general nor in relation with data analysis.

More recently, [Lehnard, 2009] has made a case for a "pragmatic concept, mode or account of understanding" (ibid., 171, 172, 182) open to the possibility that understanding can be defined implicitly in terms of "the ability it generates" (ibid, 173-174). According to such an account, understanding of a certain phenomenon or situation would occur when we have a certain ability for dealing with this phenomenon or situation in order to solve some problems concerned with it. Lenhard's example is an instance of simulation, and his major point is that in the face of simulation methods or algorithms two opposite attitudes are possible (ibid., 183): either to stick to a "traditional" concept of understanding, based on the idea that understanding is "linked to a theoretical insight" (ibid., 172), and then maintain that simulation "can provide control and options for intervention without understanding"[4] (ibid., 172), or to stick to such a pragmatic concept of understanding and then admitting that simulation models can provide a mode of understanding.

Though the notion of understanding we refer to when we emphasize lack of understanding is expressly general and weak (for the reasons indicated above), it clearly does not conform to Lenhard's pragmatic concept, and it is still well-matched with the idea that understanding requires some sort of "theoretical insight". Moreover, though it is common to consider understanding as a subjective or even psychological state, we do not want to emphasize this aspect, and rather we take the notion of understanding to differ from that of explanation not because the former is psychological and the latter logic, but rather because we consider that understanding does not necessarily involve an appreciation of the reasons for with something happens, whereas explanation does. In our sense, understanding merely happens when some structural connections characterizing the relevant phenomenon have been identified.

From the perspective we suggest, the understanding of a natural or social phenomenon is not intrinsically different from its description, at least if the latter does not simply reduce to a mere collection of data. Rather, the difference between description and understanding is a question of emphasis. The statement of a problem concerned with a particular phenomenon, and even the identification of the related problems, involves a form of description that one could take as primitive or original understanding.

[4]Though Lenhard's concern is limited to the case of simulation (which is a case we consider only indirectly), this approach is close to the one we are aiming to describe more in general and in details.

When we speak in this paper of lack of understanding of a certain phenomenon, we refer to a situation where one has not gone beyond such primitive description. Specifically, a phenomenon has been identified; a problem has been stated; and this is done by choosing a number of variables whose particular, measurable values are supposed to characterize a particular state of the phenomenon itself. However, we cannot say or establish anything else about the phenomenon, since we have not advanced any hypothesis about the mutual relation of the variables involved. In this situation, we say that the phenomenon does not admit an understanding.

From the primitive description, one can move on to a more comprehensive one, by relating different related variables to each other, or by introducing new variables, to reach a new formulation of the given problem that can help solving it. We can then take the new variables and their relations and use them to solve problems that we could not even formulate before. This is what we regard as an understanding in a proper sense, this makes us able to relate several problems and different features about the same phenomenon. This type of structural understanding is intimately related to the scale at which we consider the phenomenon and his relation between scale and understanding is so relevant that there are specific techniques in some fields (most notably in physics) to relate descriptions at different scales, as we will see in the sequel of the paper. We generally perceive some scales as being more fundamental, because our physiological experience is limited in space and time. What we need to know in these cases is how to connect our description of a phenomenon at one scale to a description of the same phenomenon at some other scale, since the fundamental quantities at different scales may be different and have different relations.

Following a notion from physics (cf. [McComb, 2008]), we term *renormalization* any transformation of a description at a fine scale J with its own set of variables in another description at a coarser scale K with fewer variables, obtained by encoding in a few new parameters the information relative to the larger number of variables that are significant at the finer scale. Hence, a renormalization of a certain description involves a (significant) reduction of the number of variables entering such a description. Since the phenomenon that is described is the same before and after the renormalization, this may possibly lead to a new understanding of the phenomenon.

As a classical example of how renormalization leads to new understanding, we note that in Newtonian mechanics, position, velocity, acceleration and mass of every particle in a system allow a complete description of this

system at any scale; as we move to very large systems like gases, understood as instances of a Newtonian system, such description is in fact not achievable in practice, especially when we try to predict the evolution of the system. Therefore, in the case of gases, we abandon the Newtonian model and find few other fundamental quantities, such as temperature and pressure, forming a basis for another description of the same phenomenon and its evolution at coarser scale. The essential point of this example is that the new description results from a renormalization of the original one connected to it by some rules that tell us how the Newtonian fundamental quantities have to be averaged to give rise the thermodynamical quantities that we can observe (at least in principle). This example can be considered the precursor of all renormalization techniques in the history of science. Renormalization provides here a macroscopic description of gases which goes together with a deeper understanding than the one that comes from their fine or coarse descriptions alone. Specifically, the connections among different scale descriptions of gases are made explicit.

Despite the crucial role that renormalization plays in the understanding of many empirical phenomena, modern data analysis does not often relies upon it. Instead, we see more and more frequently the development of fields where the maximal scale of resolution (*i.e.* the finest scale at which the phenomenon is observed) is decided on the basis of the possibility of quickly generating large amounts of data at that scale, even though it is known that a proper description of the phenomenon itself is possible at much finer scales. In this way, different descriptions at different scales are detached from each other, and we do not explain how to move from one to another. The breakdown of distinct descriptions is most often due to pragmatic needs: even though finer scale dynamics are still considered essential for a structured understanding of the phenomenon, their consideration is excluded because of a preventive decision based on inherent computational or experimental limitations. Microarrays are a fundamental case where we see this breakdown of knowledge at work [5].

[5]Note however that multiscale techniques are at the heart of modern data analysis. There are very refined signal analysis methods, such as wavelets (cf. [Mallat, 2008]), that are based exactly on idea of exploring the data at several resolution levels. But these methods are used on whatever data is available, as a way to analyze, and preprocess the data, very much in the spirit of an unstructured query of data, and they do not allow, by themselves, the type of structural understanding that we defined in this section.

3 Lack of structural Understanding: The Microarray Paradigm

In order to understand where the microarray paradigm is coming from, we need to take a short detour through some basic notions of experimental molecular biology. Recall that the DNA (deoxyribonucleic acid) molecule is made of two connected long chains (strands) of complementary building blocks (nucleotides) that encode the genetic material of the organism. Some parts of the DNA molecule are used in the cell for the production (transcription) of related messenger RNA (ribonucleic acid) molecules that are involved in the synthesis of proteins.

A DNA microarray (cf. [Hastie et al., 2000], and [Baldi and Hatfield, 2002] for a more mathematically oriented introduction) is essentially an array of microscopic sites where up to several thousands different short pieces of a single strand of DNA have been attached. Each microscopic site of the array has attached to it several copies of the same short piece of DNA strand to increase its chance of binding with the possible complementary strand of messenger RNA (mRNA). The mRNA molecules are extracted from some specific tissue and then amplified with a variety of techniques (chiefly polymerase chain reactions: cf. [Mullis et al., 1994]). They are marked with some fluorescent substance and finally dropped on each site of the microarray. Each site take a less or more intense florescence according to the amount of mRNA that binds with the strands of DNA already placed in this same site. The intensity and distribution of the fluorescence on the microarray give thus a way to evaluate the degree of complementarity of the DNA strand on the array, and the mRNA strands from the tissue.

The level of complementarity at a single site of the microarray is the *expression level* of that particular strand of DNA . The set of all expression levels on a microarray is called DNA *expression profile* of the tissue analyzed with that specific microarray.The reason such emphasis is put on the amount of different strands of mRNA is that the specific behavior of a cell depends in great part on the activity, concentration, and state of proteins in the cell itself. In turn, the distribution of proteins is influenced by the changes in levels of mRNA. This correspondence of the information on the DNA microarray with the behavior of a cell 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, a DNA microarray carries a great deal of information about cells: since thousands of strands of DNA are checked on a single microarray, one expects to obtain a description of the state of the cells from the DNA microarray. Such a description does not offer, however, an understanding of the correspondent phenomenon. Any DNA microarray supplies a particular value for a huge number of variables (the specific strand of DNA on the individual sites) and therefore displays in some way the state of the cells from the chosen tissue, but we do not know how to relate the values of the variables directly to the state of the cells. We do not use some mathematical theory to provide a systematic framework leading to the understanding of the cell behavior, starting from the variables determined by the microarray. Instead, we appeal to specific mathematical algorithms to simply classify DNA microarrays, that is, to extract a minimal, but useful, information from the huge mass of data.

If we are interested in the DNA expression profile of cancerous tissues, we begin, for a given microarray, by comparing the DNA expression profile of the cells of interest with that of some reference cells. We do this by dropping in each site of the microarray some mRNA strands derived from cancerous cells with green fluorescent pigment, and mRNA strands derived from healthy cells with red fluorescent pigment. This procedure allows to see directly in the quotient of green and red intensities which strands of DNA are activated mostly in cancerous cells and which ones in healthy cells. Regardless of the eventual role of these detected strands in the activity of the cells, we can regard their respective degree of activation as a significant and characteristic feature of these cells which distinguishes the cancerous ones from the healthy ones. We can moreover classify the cancerous cells and theirs changes in time after addition of potential drugs on the basis of their DNA expression profile. We may realize in this way which drugs are effective, or detect a common pattern that characterizes precancerous cells and so on.

Mathematics plays a crucial role in this analysis by providing the procedures that make these classifications possible. Suppose for example that we want to predict whether a patient has a specific cancer on the base of his DNA microarray expression profile X. One way to approach this problem is to turn it into a classification problem, and to consider several other patients that are known to have that specific cancer, and other patients that are known to be cancer free. Assuming that the total number of such patients is N, we measure their respective DNA expression profiles $x_1, \ldots, x_N$. The DNA expression profile of each patient is then labeled with a value y

that determines the class to which it belongs: for example, we can set this value to be $y = -1$ if the patient does not have cancer and $y = 1$ if he does have cancer. In this way, for each DNA expression profile $x_1, ..., x_N$, we get a corresponding value for y, say $y_1, \ldots, y_N$. Finally we look for a function f, belonging to a chosen space of functions, such that $y_i = f(x_i)$, $i = 1, ..., N$. Therefore, the empirical problem of prediction and classification can be reduced to a mathematical problem of approximate interpolation in which the *classifier* function f is determined, and we can predict whether the initial patient had cancer with his DNA expression profile X by computing $Y = f(X)$. If $Y = 1$, the patient will likely have cancer, and he will likely be cancer free if $Y = -1$.

The key to a successful solution of this classification problem is the identification of an appropriate space of functions $\mathcal{F}$ where we search for the best function f that satisfies $y_i = f(x_i)$, $i = 1, ..., N$. Such space has to be large enough to adapt to the diversity of data we may have, but also small enough to avoid to overfit these data: adapting f exactly to the data can entail the risk of making f susceptible on possible noise disturbances in measurements, as discussed at great length in [Ramsay and Silverman, 1997] and [Ramsay and Silverman, 2002] [6].

Because of the very large size, and apparently random nature, of DNA expression profiles, it is very difficult to determine a suitable function f in the previous classification scheme. Often the variables of the DNA expression profiles are split into groups that are significantly different, before attempting the search for the function f that solves the classification problem. This is a a technique aiming to reduce the size of the microarray variables, but it adds another layer of complexity in the mathematical analysis of the microarray that is not essential in our discussion (cf. [Kaufman and Rousseeuw, 2005]). Incidentally, this reduction of the complexity of the data was at the basis of one of the earliest attempts to find the philosophical implications of data analysis, see [Good, 1983].

The microarray technique can be used also in other biomedical applications, we have for example the important case of protein microarrays [Simpkins et al., 2004], that directly check the activity level of several proteins at once. The two basic conditions that have to be satisfied to make microarray possible are: the ability to build large and inexpensive arrays of testing

[6]The structure of the general classification procedure we outlined here will be used again in the description of the data analysis methods of Sections 4.1 and 4.3.

sites in a reasonably short time; and the ability to detect, in a reliable way, the expression level of the specific substances that are being probed at each site. Solutions to biological and medical questions may be based on the properties of microarrays, even though we know microarrays do not provide any explanation of the biological functions, at fine scales, that determine the macroscopic property—for example, presence or absence of disease—we are interested in. The coarse scale biology, namely the presence of cancer in an individual, is thus detached from the fine scale microbiology, namely the specific structure of genes interactions, even though we know that some fundamental piece of information related to our problem lies at the fine scale level. It is in this sense that there is already a breakdown of structural understanding, before we even start to analyze the data. And the robustness and legitimacy of the solutions gathered from microarrays is often extrapolated by artificially adding noise to the data, and testing the reproducibility of the classification results on these perturbed data (see [Simon et al., 2003], section 9.5 and [McShane et al., 2002].), rather than by probing the possible biological reasons that could explain the solutions we found[7].

* * *

To summarize, the mathematical study of microarrays is a clear example of prediction and inference from unstructured data that is a trademark of modern data analysis. According to a more traditional paradigm of mathematization of empirical sciences, especially exemplified in Newtonian science, mathematical techniques are used in order to establish a conceptual framework which provides a technical characterization of the phenomenon to be studied. This characterization is understood as a mathematical structure whose intrinsic relations make possible to associate a formalism to the phenomenon itself. This structure provides an objective representation of the phenomenon that makes it intelligible by involving appropriate interpretative categories (like, speed, force, energy, etc.)[8]. Moreover, the other purpose of the mathematical structure is exactly to transform the phenomenon in a

[7]This is the case especially for large microarrays, like DNA microarrays. The analysis of smaller microarrays, such as protein microarrays, is often followed by *in vitro* validation of the results inferred through the microarray.

[8]Though in a quite different context (that of population genetics), M. Morrisson ([Morrison, 2006], p. 340) has insisted on this point, by arguing that "mathematical abstraction can play an important role in *shaping* the way we think about and hence understand certain phenomena".

piece of mathematics, with local empirical inputs, to make possible to rely on known techniques in order to prove theorems, solve problems and compute values of quantities related to the phenomenon. In this way mathematics works together with a conceptual interpretation of the phenomenon, which is the modern form of what Aristotle, in the beginning of physics, termed $\delta\iota\alpha\acute{\iota}\rho\varepsilon\sigma\iota\varsigma$: the reduction of the sensible data to a system of general principles.

Whereas in the traditional paradigm the identification of invariants is largely a conceptual endeavor, in the case of the microarray paradigm, it is essentially the outcome of an algorithm. In the main examples of Section 4 we will see how the data arising from a phenomenon can be directly used to solve specific, narrowly defined, problems, making superfluous the conceptual identification of the variables that come with an understanding of the phenomenon at different scale levels [9].

Needless to say, the data used in data analysis methodologies can hardly count as raw (cf. [Harris, 2003]). As we have already emphasized, in order for data analysis to apply, a phenomenon must be identified and a problem about it must be stated. This is done by measuring values for a large number of variables and different states of the phenomenon are taken to depend on the measured data. In our parlance, this is a description of the phenomenon itself, and this description is not only necessary for collecting data, but is also already sufficient for organizing them. Still, there is a difference between large amount of data, organized according to a certain description of a given phenomenon (which assumes that the data are related to this phenomenon and to particular problems relative to it), and a "model of data" in Suppes' classical sense [Suppes, 1962]. According to Suppes, models of data belong to a "hierarchy of model that connect data to theory" ([Harris, 2003], 1508). Further, they are "designed to incorporate all the information about the experiment which can be used in statistical tests of the adequacy of the theory" ([Suppes, 1962], 258). Hence, models of data strictly depend on a theory, and–however it might be conceived–a theory cannot help depending on some sort of understanding, which is just what is lacking, in the case of microarrays.

There is thus something in between a body of raw data and a model of

[9]The study of microarrays show that mathematics can be used to identify some useful characteristics in a huge amount of data before, and independently of, any conceptual interpretation. Even though the mathematical structure of the classification method used for DNA microarrays does not fully capture the way the microarray *paradigm* can be used as a methodological guiding principle in data analysis.

data, and it is here that the microarray paradigm applies. In cases like these, data are chosen and measured according to a basic description of a certain phenomenon, but no theory and/or understanding is available for transforming them in a model in Suppes' sense. This does not mean of course that no theory and/or understanding are necessary in order to provide the relevant data. In the very case of microarrays, a significant background of genetic biology, histology, and chemistry, is necessary to build such an array, and these tools come with a very sophisticate understanding of related phenomena. Still, these theories and the relative understanding are not illuminating the phenomenon and problem to be studied, and therefore in this sense a microarray is not a model of data for them. The microarray paradigm prescribes, in cases like these, to apply mathematics (usually appropriate computational algorithms) not in order to get a reproducible curve that fits available data in an appropriate space (of the right dimension), or to transform them in some suitable way so to produce a model ([Harris, 2003], 1510-1512), but rather in order to question the data as such, to solve the problem before any understanding [10]. The way this is done depends of course on the particular case under examination, and the microarray paradigm can prescribe neither a particular method nor a family of methods to do it To understand how the microarray paradigm can apply in other cases, essentially different from the very case of microarrays (both for the nature of the phenomenon and problem under examination, and the mathematical theories involved), other examples are needed. We hope they might be useful to clarify the way in which the microarray paradigm is exemplified in a variety of fields.

4 The microarray paradigm at work

Our selection of data analysis methodologies that exemplify the microarray paradigm is very limited with respect to the great variety of methods that we could explore. We have decided to highlight a few where we can see very clearly that no understanding on the phenomenon is gained while solving the problem.

[10]The very construction of a model (of data or of phenomena) requires some understanding and we purposely chose not to address the longstanding question of whether a model represent reality, or is instead only an instrument for prediction. Our objective is to show something different and stronger, i.e. that it is possible to make predictions without models, and therefore without understanding.

The first method that we describe in Section 4.1 makes use of artificial neural networks, a well established technique that was first developed in 1943 in [McCulloch and Pitts, 1943]. We focus on an applications of neural networks to hurricane strength prediction, since this setting was already touched upon in Section 2, and since it demonstrates that interpretable simulations can be less effective than non-interpretable techniques. We move then in section 4.2 to a subtle use of mathematics in the study of physiological data with techniques from nonlinear dynamics. What is remarkable in this application is that it is necessary to invoke quite advanced theorems on the immersion (embedding) of geometrical structures in large ambient spaces. And yet the actual model of the physiological data remains hidden up to the end. Finally, section 4.3 summarizes one of the most interesting techniques in data analysis, boosting, that purposely does not build interpretable, efficient models, and trades them for a powerful combination of weak, not-interpretable ones.

Two more methodologies that fit within the approach of the microarray paradigm are briefly described in Section 5, but with a different purpose: we will show that the lack of structural understanding in data analysis does not diminish the use of sophisticated mathematical ideas, but on the contrary it increases their relevance to very concrete problems.

4.1 Neural networks and hurricanes forecasting

We highlight in this subsection a method that uses neural networks to forecast the intensity of winds of hurricanes (up to 48 hours in advance), on the basis of a set of available meteorological data (cf. [Baik and Paek, 2000]).

We recast here neural networks in the language of classification problems as in Section 3. According to this viewpoint, neural networks are nothing more than a a very specific choice of space of functions of the input variables used to perform approximate interpolation (see [Hastie et al., 2001], chapter 11 for more on this interpretation of neural network in terms of approximation in functional spaces).The structure of the functions in this functional space provides a primitive modelization of the dynamics of actual neurons.

The construction of neural networks can be seen as a two steps process. First, input variables $X = (X_1, ..., X_N)$ are preprocessed to get new variables $Z_i = \sigma(\alpha_{0i} + \sum_j \alpha_{ij} X_j)$, $i = 1, ..., M$, where the parameters α_{0i} and α_{ij} are to be determined, and where the function σ is selected to mimic an important property of the activation patterns of neurons, i.e. the fact that their firing is very small when the input is small, and then suddenly large

when the stimulus of the neurons is above a set threshold. Choosing σ to be $\sigma(r) = \frac{1}{1+e^{-r}}$ provides exactly such sudden switch of the intensity of $\sigma(r)$ as r goes from large negative values (when $\sigma(r) \approx 0$) to large positive values (when $\sigma(r) \approx 1$)[11]. Second, after defining the variables Z_i, we try to express the output variable Y as a linear combination of the variables Z_i, i.e. we adjust parameters β_0 and β_i so that $Y \approx \beta_0 + \sum_i \beta_i Z_i$. This last step is essentially the approximate interpolation step that we introduced for the classification of DNA expression profiles in Section 3, with the notable difference that the output variable is allowed to assume any real value.

It is remarkable that (cf. [Baik and Paek, 2000]) the use of neural networks gives better forecasts of the intensity of winds than the best available simulations of atmospheric dynamics. In this context, the input variables X_i are sets of measured meteorological data relative to a developing hurricane, and the output variable Y is the intensity of winds of the hurricane after 48 hours. The crucial point of this method is that the structure of neural networks does not express any understanding of the hurricane dynamics. It does not mirror in any understandable way the structure of the atmosphere: the specific problem is solved, but with no new knowledge of the phenomenon. Note moreover that only the ability to access a large quantity of measurements for the hurricane during its development allows this technique to work, in line with the general tenants of the microarray paradigm.

An interesting consequence of such opaque way to make predictions is the compelling need of validating them very carefully. In [Kalnay, 2003], individual predictions are validated by looking at *ensemble* weather predictions, where several predictions are made, possibly with different methodologies and different initial conditions, and only the most likely result, or group of results, is then used for the actual forecasting. This technique clearly does not improve basic understanding since, even though we may trust more the prediction of a collection of methods because of a well understood a posteriori statistical analysis, the individual methods will not be more transparent because of this postprocessing. There is a similarity between the ensemble forecasting and the boosting technique described in section 4.3, these paral-

[11] The historical reason for defining the variables Z_i in the neural network procedure was the belief that a plausible approximation of the structure of interconnected neurons in the brain may be useful to approach complex problems (see the discussion of [Bailer-Jones and Bailer-Jones, 2002] on the analogy to the brain). The distinctive feature that is believed to be the key for the success of neural networks is their ability to adapt to input data in a non-linear manner (cf. [Ripley, 1996]).

lels are one more sign of a convergence, in different fields, to similar weak approaches.

4.2 Data-driven control of seizures

Another very interesting problem in which a lot can be done without understanding the details of the phenomenon is the detection and control of abnormal physiological behavior. Both electric cardiogram (ECG) and electric encephalogram (EEG) show complex, but deterministic, behavior at the onset of heart arrhythmia (cf. [Christini et al., 2001]) and seizures, respectively (cf. [So et al., 1998]). This behavior can be described with relatively simple systems of differential equations $\dot{x}(t) = S(x(t))$, where $x(t) = [x_1(t), ..., x_n(t)] \in \mathbb{R}^n$ and $\dot{x}(t)$ is the vector of the derivatives of each variable (cf.[Kantz and Schreiber, 2003]).

If the system of differential equations admits a low-dimensional region $\mathcal{G}$ contained in $\mathbb{R}^n$ where most trajectories $x(t)$ that satisfy $\dot{x}(t) = S(x(t))$ converge, we say that the system of differential equations has a low dimensional *attractor* and knowledge of $\mathcal{G}$ encodes some important features of solutions of $\dot{x}(t) = S(x(t))$. The question is then whether $\mathcal{G}$ can be inferred from measurements of the trajectories $x(t)$.

In particular, assume now that we measure a single quantity $q(t)$ derived from the trajectories components, say $q(t) = G(x_1(t), ..., x_n(t))$, where G is a differentiable function. We measure $q(t)$ at a uniformly sampled set of time values $t = t_0, t_0 + dt, ..., T$, with dt a small sample unit, $T = t_0 + Qdt$ and Q some large integer value. These discrete measurements of $q(t)$ can be used, if the number of samples Q is sufficiently large, to gain a rough geometrical understanding of the system $\dot{x} = S(x)$, through the use of so called delay maps. Essentially a delay map takes a fixed number d of consecutive points in the measurement $q(t)$ at constant intervals of τ starting from a given point $q(\bar{t})$ and maps them in the space R^d. If d is large enough the set of all points mapped to $\mathbb{R}^d$ from a single measured quantity $q(t)$ (for all possible values of $\bar{t}$), will often look like a deformation of the attractor $\mathcal{G}$ where the trajectories of the system $\dot{x} = S(x)$ converge (cf. [Alligood et al., 1996], [Takens, 1981], and [Sauer et al., 1991]). It is somewhat surprising to realize that even measuring a single trajectory variable can be sufficient for this technique to work, i.e. we can take $q(t) = x_{i_0}$ for some index i_0.

In the case of EEG data, delay maps allow, at least in principle, a data-driven control of seizures as outlined in [So et al., 1998], see also [Kapitaniak,

1996] and [Ott et al., 1990]. The key idea is to reconstruct the attractor $\mathcal{G}$ associated to the EEG data from the delay map and to use this information to prevent the trajectories from displaying the deterministic behavior that is often a precursor to seizures, by using a suitable, slight perturbation of the system. Since we can get an idea of the important features of the dynamics from the delay map, we do not need a previous knowledge of the form of the system, as long as we assume a simple enough deterministic dynamics before a seizure. The information deduced from the delay map is enough to perform some type of control of the seizure and, ideally, to avoid it.

What is essential to the functioning of this technique is the ability to access long measurements of EEG data, which allows a rough geometrical representation of the salient features of the dynamical evolution of the neurons whose activity generated the seizure captured by the EEG. So the exact equations in $\dot{x} = S(x)$ are not necessary, and a graphical description of the underlying dynamics is sufficient for this control technique to work. Note again that, in principle (cf. [Kantz and Schreiber, 2003]), even a single long measurement, i.e a measurement followed for a sufficiently long time, can lead to a geometrical representation of the dynamics of the underlying system. This implies that under suitable conditions on the dynamics of the system, almost any single long measurement can be used to infer the general characteristics of the system.

We can understand this result in view of the microarray approach if we see that the diversity of measurements that is asked for in the microarray approach is provided by the number Q of time samples of the measured quantity $q(t)$. This requirement is sufficient in this case because the dynamics of the measurements is so complex, that the longer we record them in time, the more we learn about the whole system.

Note also that the dynamics of EEG measurements cannot be equated to the dynamics of single neurons. In practice the electrical signals we measure will be the sum of many electrical signals from nearby neurons. This is also the fundamental reason why the actual dynamics of the neurons is not directly used for the control of their collective behavior, i.e. of global phenomena such as seizures. These techniques have been used in practice with success for small groups of neurons, while their applicability to large systems is still the subject of active research.

4.3 Boosting

The previous two examples show to a great extent the characteristics that we expect from an approach based on the microarray paradigm. Let us now consider a mathematical technique which is having a consistent impact on data analysis since its domain of applicability is wide, but whose ability to shed light on the structure of the phenomenon is quite limited. This method goes under the name of *boosting*, and it is remarkably well suited to be the backbone of a microarray-paradigm-based data analysis. Unlike what we have done in the previous two subsections, here we only describe the algorithmic outcome of this method, without specific applications. However, it turns out that boosting is a way to put together weak classifiers (*i.e.* algorithms that can distinguish classes of objects just a bit better than randomly) to obtain an arbitrarily strong classifier, i.e. an algorithm that classifies objects correctly most of the time. All the empirical classification problems we described earlier in the paper could in principle benefit from this technique. We give a quick description of boosting that closely follows [Hastie et al., 2001], chapter 10.

Suppose we have a standard classification problem and that we consider only two classes of phenomena, each labelled by a variable y that can take values in the discrete set $\{-1, 1\}$. Suppose also that we can measure some variables x associated with the phenomena belonging to both classes. As we have seen in Section 3, a classifier is a function f such that $f(x) = 1$ if x is a measurement of a phenomenon from class $\{y = 1\}$ and $f(x) = -1$ if x is a measurement from class $\{y = -1\}$. We take a weak classifier to be a function g that correctly classifies measurements x with frequency $f = \frac{1}{2} + \epsilon$, with ϵ fairly small, *i.e.* the frequency of right guesses is only slightly better than a random guessing. Recall now that the function f is adjusted (trained) according to a set of training measurements $\mathcal{X} = \{x_1, \ldots, x_n\}$ where each element of the set is known to belong to one of the two classes. The key advance of boosting methods is to generate slightly different versions of the classifier f, say f_1,...,f_n, by modifying the importance of each individual measurement in $\mathcal{X}$, before training the classifier. This is done according to a rather sophisticated strategy that we do not report here (cf. [Hastie et al., 2001] chapter. 10). It suffice to say that the final outcome of the boosting algorithm is a classifier $\bar{f}(x) = sign[\sum \alpha_i f_i(x)]$, where the coefficients α_i determine the importance of each individual weak classifier in the overall classifier defined by $\bar{f}$. Note that $\bar{f}(x)$ can only take values in $\{-1, 1\}$ and it

gives therefore a formally correct prediction on the class of belonging of x. It has been observed in some experiments that if the rate of correct classification of each individual weak classifier f_i is just 54%, then boosting can give a final $\bar{f}$ with success rate of 88% (cf. [Hastie et al., 2001] chapter. 10).

In [Freund and Schapire, 1999] (see page 10 of the English translation), the initial developers of the boosting technique point out that such a technique depends on "a shift in mind set for the learning-systems designer: instead of trying to design a learning algorithm that is accurate [...], we can instead focus on finding weak learning algorithms that only need to be better than random." This approach to data analysis fits perfectly in the context of the microarray approach, as it suggests that weak techniques can be combined to obtain strong techniques that are devoid of any meaningful interpretation.

5 Discussion. What Role for Mathematics?

Our claim in this paper is that modern data analysis, faced with increasingly data-heavy problems, has challenged mathematics to assume a different role in its relationship to phenomena. We have argued that, unlike the traditional approach in which mathematics models a phenomenon while fostering its understanding, the modern paradigm used in data analysis gives up understanding, in favor of increasingly powerful forecasting tools. We have summed up our claim under what we called the microarray paradigm, which insists on the fact that enough, diverse data may help to solve most questions related to a specific phenomenon, even though they may not shed any light on its actual understanding. Our choice of terminology is heavily influenced by the remarkable success that DNA microarray technology has had in modern biomedical applications, but we have included in our paper a series of examples from other areas of current research, to show that, in fact, problems far removed from microbiology lead to methods that can be interpreted and systematized under the heading of the microarray paradigm.

We discuss now two recent works that tried to understand the methodological and philosophical implications of data analysis, and that come close to our viewpoint. In section 3.1 of [Bailer-Jones and Bailer-Jones, 2002], Daniela and Coryn Bailer-Jones contrast "data analysis models" with "theoretical scientific" ones. In elaborating the latter models, they argue that "one is interested in determining the values of some important, physically mean-

ingful parameter or parameters with the aim of better understanding" the relevant phenomenon (they consider the example of a model for the temperature variation in the Earth's atmosphere). In data analysis, and especially in computational data analysis, instead, "the primary interest is in training a model to make predictions" and "the model parameters are a mean to an end and not necessarily of physical significance themselves". It follows from their considerations that data analysis models "are conceived without regard to the theories and concepts of the various problems to which they can be applied" and "are sufficiently general so that they can be applied to a wide class of problems with only general requirements having to be met". As a consequence, "data analysis techniques are not specific to the type of data that are modelled" and "are designed to be independent of specific applications" or "application-neutral". It seems thus that, for D. and C. Bailer-Jones, data analysis is simply a set of techniques for computation and classification, and it is, in fact, disjoint or poorly related to the study of phenomena. To the useful techniques of traditional mathematization, data analysis is adding new ones that are neutral, and that are used in practice more by relying on the analogy among different phenomena than by developing insight in the specific phenomenon.

While this *neutrality* viewpoint certainly captures some important characteristics of data analysis, we note that any mathematical technique is neutral by itself, and the use of analogy in shifting the same mathematical structure from one field to another has been extremely successful also in a more traditional approach to science. Examples are the use of techniques from quantum field theories in condensed matter physics [Altland and Simons, 2006], and the use of techniques from the theory of spin glasses to optimization and image processing problems [Nishimori, 1999]. We believe that data analysis is indeed an innovative way to study phenomena, and not merely a collection of convenient techniques. And we have tried to describe in this paper the uniqueness and novelty of this approach to scientific discovery.

The author that explored most carefully the impact of computational methods in science is Paul Humphreys. His work, culminating in [Humphreys, 2004], is mainly concerned with the way mathematical models of physics are actually transformed into computational models with adjustable parameters, or "computational templates", as Humphreys calls them, which have effective predictive power. According to him, the computational models cannot be developed unless the mathematical models have been worked out beforehand. This means that a certain degree of understanding is necessary for

computational science to work, even if this understanding is then weakened in the development of effective predictive templates. In [Humphreys, 1995] Humphreys argues that the effective prediction power of the latter is reached through the use of a few mathematical forms that are valid across disciplines (similarly to the neutral techniques of [Bailer-Jones and Bailer-Jones, 2002]).

This is certainly true in many cases. Humphreys mentions, as an example, the case of differential equations, some of which famously apply to disparate fields. Still, he does not highlight enough the profound shift in perspective that leads often to radical, not simply pragmatic, removal of understanding. In [Humphreys, 2004] (section 1.2 pp. 6-8), he argues that "our own intellectual and computational capabilities as human beings is no more the benchmark of scientific thought", and in [Humphreys, 2009] (section 2) he adds that "Computational science introduces new issues into the philosophy of science because it uses methods that push humans away from the center of the epistemological enterprise" and that:"[...]the situation within which humans deal with science that is carried out at least in part by machines [is] the hybrid scenario and the more extreme situation of a completely automated science [is] the automated scenario [...]". In our view the changes brought by the methods of data analysis are not simply an issue of automated versus human science. They depend, much more fundamentally, on they way a phenomenon is approached, namely on the fact that the aim of solving a specific question and getting predictive power comes first, and is often opposed to, the effort of getting any understanding.

A key question is rather whether the new modality of interaction of mathematics and empirical science can be fruitful not only in solving problems, but in fostering new ideas in mathematics as well. While the answer may be negative at some level, it opens a new perspective on the prominence of mathematical thought in science. Mathematics has often benefited from ideas and formalisms developed in physics, where a less rigorous development of mathematical concepts pairs with the profound intuition that is generated by the structural understanding of phenomena.

The lack of this structural understanding in data analysis does not diminish the use of sophisticated mathematical ideas, but on the contrary it increases their relevance to very concrete problems. However, data analysis cannot be a source of ideas and methods the way physics has been and continues to be. The microarray paradigm affects the development of methodologies by liberating mathematics from the necessity of strong isomorphisms with phenomena. Mathematical ideas, no more constrained by the actual

structure of the phenomenon, are free to directly affect the way we solve problems. This process realizes itself not only by a strong use of analogies, but literally by forcing the data to fit into known mathematical structures of sufficient complexity.

Take the case of *chemical graph theory* (see [Bonchev and Rouvray, 1991]). In this field, the graph of the links of the atoms in molecules is considered crucial for the prediction of specific chemical properties of molecules, such as toxicity, or carcinogen effect. Only the configuration of the links is considered, not the angles or the distances among atoms, so that we are effectively looking at the topology of the graph. Prediction of a given chemical property for a target molecule is obtained in two steps: first, by searching for topological invariants that assume very close values for the graphs of a large number of molecules with the chemical property; second, by computing the topological invariant for the target molecule, and checking whether it assumes a value close to the cluster of all the values computed in the first step. Because it may not be transparent how the most suitable invariant relates to a chemical property, this application of graph theory is very much in the spirit of the microarray paradigm. At the same time, the topological invariants have to be searched among those that carry significant graph properties, and therefore we need a sophisticated knowledge of graphs simply to be able to propose new invariants to test in the basic procedure highlighted above. Instead of being a source of discovery of new topological invariants for graphs, chemistry becomes, in the context of this application, only a recipient of ideas from advanced graph theory.

Another instance of forcing of mathematics on the data is the use of geometry in text organization and labeling. Here the assumption is that it is useful to derive a well defined geometrical manifold from the data of a problem, because on such geometrical object it is possible to define functions that can be used for further manipulation of the object. Consider, following the review in [Coifman and Maggioni, 2008], a collection of articles from a multidisciplinary scientific journal. If we fix a large set of N, randomly chosen words, the frequency of the n-th word for each article can be used to define the n-th coordinate of a point in a N dimensional space. We can then associate this point to the article itself. We obtain in this way a cloud of points in the N-dimensional space, where each point is associated to an article. There are standard techniques (cf. [Ramsay and Silverman, 1997]) to identify the directions of maximal variance of the cloud of points, and these main directions can be used to visualize a low dimensional image of

the set of points. The geometry of the low dimensional image carries interesting information on the similarity of different articles, and even of different topics; for example, it is possible to find out that earth science articles are very close geometrically to biology articles, and, not surprisingly, mathematics and physics articles are close to each other. Conversely, the shape of the geometrical image can provide automatically the labeling of article in distinct topics. The body of work reviewed in [Coifman and Maggioni, 2008] (see also [Szlam et al., 2008]) makes the further subtle observation that there are suitable functions on the geometrical manifold that can be used to preserve the edges and the boundaries of the subsets associated to each topic. These functions, too involved to be described here, are useful when we try to go from a small set of labeled articles to a larger set of unlabeled ones. This process of propagation of labels can be seen as a diffusion along the geometrical image of the set of articles, and the functions defined on the object make diffusion across edges difficult, so that mislabeling of unlabeled articles is minimized. The resulting field of *diffusion geometry* (cf. [Szlam et al., 2008]) has a very strong emphasis on finding techniques that lead from unstructured large discrete sets, to geometrical, smooth objects, that are much better understood mathematically. The emphasis on large data sets is clearly reminiscent of the microarray paradigm, and the effort put in leading the data to a target field of mathematics, exclusively for manipulation and analysis purposes, is exactly what we mean by forcing.

These examples should clarify why the effectiveness of mathematics in relating to phenomena is not affected by the methods of modern data analysis. Only the flow of ideas is somewhat inverted, as mathematics is required to take the lead in providing the setting and the quantities necessary for solving problems, without receiving much insight in the process.

Mathematics becomes perhaps the only domain in which to develop structural understanding, since such pretense is lost in the study of phenomena. Ideas are then forced upon the phenomenon in problem solving, only temporary, and with little expectations that go further than the solution of the problem. Scientific methods may become weak, but the mathematical language in which they are phrased will be increasingly complex, as we attempt to mould our desires, coarsely, upon reality.

References

P. Achinstein. *The Nature of Explanation*. Oxford University Press, New York, 1983.

K. T. Alligood, T. Sauer, and J. Yorke. *Chaos. An introduction to Dynamical systems*. Springer, New York, 1996.

A. Altland and B. Simons. *Condensed matter field theory*. Cambridge University Press, Cambridge, 2006.

J. Baik and J. Paek. A neural network model for predicting typhoon intensity. *J. Meteor. Soc. Japan*, 78:857–869, 2000.

D. M. Bailer-Jones and C. A. L. Bailer-Jones. Modeling data: Analogies in neural networks, simulated annealing and genetic algorithms. In L. Magnani, editor, *Model-Based Reasoning: Science,Technology, Values*, pages 147–165. Kluwer-Academic, Dordrecht, 2002.

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.

R. Batterman. *The devil in the details*. Oxford University Press, Oxford, 2002.

R. Batterman. On the explanatory role of mathematics in empirical science. *British Journal for the Philosophy of Science*, to appear.

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

D. J. Christini, K. M. Stein, M. S. Markowitz, S. Mittal, D. J. Slotwiner, M. A. Scheiner, S. Iwai, and B. B. Lerman. Nonlinear-dynamical arrhythmia control in hyumans. *Proceedings of the National Academy of Science*, 98:5827–5832, 2001.

A. Cichocki and S. Amari. *Adaptive Blind Signal and Image Processing*. Wiley, New York, 2002.

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

D. Donoho. High-dimensional data analysis. the curses and blessings of dimensionality. 2000. AMS Lecture, Math. Challanges of the 21st Century, 2000. Available at www-stat.stanford.edu/ donoho/Lectures/AMS2000/Curses.pdf.

Y. Freund and R. Schapire. A short introduction to boosting. *J. Japan. Soc. for Artifcial Intelligence*, 14(5):771–780, 1999. In Japanese, English translation available at www.cs.princeton.edu/ schapire/boost.html.

M. Friedman. Explanation and scientific understanding. *The Journal of Philosophy*, 71:5–19, 1974.

I. J. Good. The philosophy of exploratory data analysis. *Philosophy of Science*, 50(2):283–295, 1983.

T. Harris. Data models and the acquisition and manipulation of data. *Philosophy of Science*, 70(5):1508–1517, 2003. Proceedings of the 2002 Biennial Meeting of the Philosophy of Science Association. Part I: Contributed Papers.

T. Hastie, R. Tibshirami, M. B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W. C. Chan, D. Botstein, and P. Brown. 'gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. 2000. Available at: genomebiology.com/2000/1/2/research/0003/.

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

P. Humphreys. Computational science and scientific method. *Minds and Machines*, 95(5(4)):499–512, 1995.

P. Humphreys. *Extending ourselves*. Oxford University Press, New York, 2004.

P. Humphreys. The philosophical novelty of computer simulation methods. *Synthese*, 169(3):615–626, 2009.

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

H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis*. Cambridge University Press, Cambridge; New York, 2003.

T. Kapitaniak. *Controlling Chaos*. Academic Press, Boston, 1996.

L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience, New York, 2005.

T. W. Lee. *Independent Component Analysis. Theory and Applications*. Kluwer A.P., Boston, 1998.

J. Lehnard. The great deluge: Simulation modeling and scientific understanding. In H. W. de Regt, S. Leonelli, and K. Eigner, editors, *Scientific Understanding: Philosophical Perspectives*, pages 169–186, Pittsburgh, 2009. University of Pittsburgh Press.

S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, San Diego, 2008.

W. D. McComb. *Renormalization Methods: A Guide For Beginners*. Oxford University Press, New York, 2008.

W. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 7:115–133, 1943.

L. M. McShane, M.D. Radmacher, B. Freidlin, R. Yu, M. Li, and R. Simon. Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics*, 18:1462–1469, 2002.

M. Morrison. Scientific understanding and mathematical abstraction. *Philosophia*, 34:337–353, 2006.

K. B. Mullis, F. Ferre, and R. A. Gibbs. *The Polymerase Chain Reaction*. Birkhauser, Boston, 1994.

H. Nishimori. *Statistical Physics of Spin Glasses and Information Procesing*. Oxford university press, New York, 1999.

E. Ott, C. Grebogi, and J. A. Yorke. Controlling chaos. *Phys. Rev. Lett.*, 64:1196–1199, 1990.

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

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

H. W. De Regt and D. Dieks. A contextual approach to scientific understanding. *Synthese*, 144:137–170, 2005.

B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, New York, 1996.

W. Salmon and P. Kitcher. *Scientifical Explanation*, volume XIII of *Minnesota Studies in the Philosophy of Science*. University of Minnesota Press, Minneapolis, 1989.

T. Sauer, J. Yorke, and M. Casdagli. Embedology. *Journal of Statistical Physics*, 65:579–616, 1991.

M. Scriven. Explanations, predictions, and laws. In H. Feigl and G. Maxwell, editors, *Scientific Explanation, Space, and Time*, volume 3 of *Minnesota Studies in the Philosophy of Science*, pages 170–230. University of Minnesota Press, Minneapolis, 1962.

R. M. Simon, E. L. Korn, L. M. McShane, M. D. Radmacher, G. W. Wright, and Y. Zhao. *Design and Analysis of DNA Microarray Investigations*. Springer, New York, 2003.

F. Simpkins, E. Kohn, V. Espina, A. F. Petricoin III, and L. A. Liotta. Beyond genomics to functional proteomics. *Genomics and Protenomics*, 4 (9):S7–S14, 2004.

P. So, J. T. Francis, T. I. Netoff, B. J. Gluckman, and S. J. Schiff. Periodic orbits: A new language for neuronal dynamics. *Proceedings of the National Academy of Science*, 74:2776–2785, 1998.

P. Suppes. Models of data. In E. Nagel, P. Suppes, and A. Taski, editors, *Logic, Methodology and Philosophy of Science*, volume 898 of *Lecture Notes in Mathematics*, pages 252–261. Stanford Univ. Press, Stanford, 1962.

A. D. Szlam, M. Maggioni, and R. R. Coifman. Regularization on graphs with function-adapted diffusion processes. *Journal of Machine Learning Research*, 9:17111739, 2008.

F. Takens. Detecting strange attractors in turbolence. volume 898 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, New York, 1981.

S. Toulmin. *Foresight and Understanding*. Harper and Row, New York, New York, 1963.

U. Trottenberg, C. W. Oosterlee, and A. Schuller. *Multigrid*. Academic Press, San Diego, 2000.

J. D. Trout. Scientific explanation and the sense of understanding. *Philosophy of Science*, 69:212–233, 2002.

E. Weber. Explaining, understanding and scientific theories. *Erkenntnis*, 44: 1–23, 1996.