Agnostic Science. Towards a Philosophy of Data Analysis

D. Napoletani · M. Panza · D. C. Struppa

Published online: 20 June 2010
© Springer Science+Business Media B.V. 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

D. Napoletani (✉)
Department of Mathematical Sciences, George Mason University, Fairfax, VA 20030, USA
e-mail: dnapolet@gmu.edu

M. Panza
IHPST, CNRS, Univ. Paris 1 and ENS Paris, Paris, France
e-mail: Marco.Panza@univ-paris1.fr

D. C. Struppa
Department of Mathematics and Computer Science, Chapman University, Orange, CA 92866, USA
e-mail: struppa@chapman.edu
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 Sect. 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 Sect. 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