Ann Ist Super Sanità 2022 | Vol. 58, No. 3: 151-153
DOI: 10.4415/ANN_22_03_01

EDITORIAL
The challenge of complexity in the Big Data era: how to ride the wave of high-dimensional data revolution

Cecilia Bossa, Igor Branchi, Barbara Caccia, Evaristo Cisbani, Carla Daniele, Giuseppe D'Avenio, Giuseppe Esposito, Francesco Facchiano, Gianluca Frustagli, Roberta Valentina Gagliardi, Andrea Galluzzi, Daniele Giansanti, Guido Gigante, Alessandro Giuliani, Loredana Le Pera, Maurizio Mattia, Sandra Morelli, Ornella Moro, Alessandra Palma, Antonio Pazienti, Orietta Picconi, Elisabetta Pizzi, Cecilia Poli, Irene Ruspantini, Sabrina Tait and Olga Tcheremenskaia for the Complex Systems and Data Science Group

Istituto Superiore di Sanità, Rome, Italy

A famous joke, reported in [1], clarifies some epistemological fault lines dividing different scientific traditions.

A very rich man, very fond of horse races, hired a top-class mathematician (e.g., Kurt Godel) and a top-class physicist (e.g., Albert Einstein) to build a model enabling him to exactly predict the winner of any horse race. After one year, both scientists returned to the rich man with their results. Godel said "Sir, I cannot say which is the specific horse who will win the race, but I discovered that the solution to the problem exists and it is unique".

The sponsor is not satisfied at all and asks Einstein if he can say something more practical and useful, Albert says "Why did you ask Kurt? You should know mathematicians have no sense of reality; on the contrary, I have the exact solution indicating the specific winner of the race. It applies only in the case of spherical horses but I am convinced this is not a problem and in any case it will be surely solved by inserting some minor adjustable parameters into the model".

Beside the delusion of the rich man, the joke reports two crucial limits of the classical mathematical and physical way of reasoning when dealing with biology: the lack of interest of both too abstract solutions and ideal cases to approximate real world. Notably, the rich man should be equally disappointed by a very long list of the "statistically significant features" differentiating frequent winners from less performant horses in the last two centuries proudly proposed to the sponsor by a famous geneticist with the help of a bioinformatics team.

We will go back in the following to this addition to the joke because it has to do with a new "player" of the game: the "machine intelligence" approach that mixes up the cards.

In summary, the joke reminds us we need an integration of different sciences to overcome the lack of real innovation [2] of nowadays research work. This process of integration is actually on the run [3] and we are already part of it. In this context, the revitalizing of the time-honoured science integration tradition of our Institute (where it is still possible to meet physicians and biologists involved in statistical epidemiology and multidimensional data analysis or physicists participating to neuroscience projects) is one of the reasons that fostered the creation of our group.

This is by no means an isolated initiative: the same urgent need generated many interdisciplinary groups in different Research Institutes all around the world. Just to name a few: the "Emergent Dynamical Systems Group" in New York (https://www.science.org/doi/full/10.1126/sciadv.aat1293) the "Complexity Science Institute" in Potsdam (https://www.pik-potsdam.de/en/institute/departments/complexity-science), the "Theoretical and Scientific Data Science Group" at SISSA in Trieste (https://datascience.sissa.it/).

Beside the peculiarities of these groups: some more focused on applicative studies, some more theoretically oriented, some very informal (like ours), some more academically structured, they all share the need of integration of different fields of enquiry to face the new challenges that cannot be faced by single scientific traditions.

To better focus the "state of the art" of the relations between biomedical and more quantitatively oriented traditions, we need to make a short digression toward a better understanding of the concept of complexity.

One of the fathers of information science, Warren Weaver, in his fundamental "Science and Complexity" 1948 paper [4], proposed a three-class partition of sci-

Address for correspondence: Alessandro Giuliani, Dipartimento Ambiente e Salute, Istituto Superiore di Sanità, Viale Regina Elena 299, 00161 Rome, Italy. E-mail: alessandro.giuliani@iss.it.

ence into: 1) Organized Simplicity, 2) Disorganized Complexity, and 3) Organized Complexity.

The first class (Organized Simplicity) refers to the classical use of quantitative methods in science. Class 1 problems permit an extreme abstraction (e.g., a planet can be considered as a dimensionless "material point": we are into a "spherical horse" approach that perfectly works in many situations). This approach allows to generate differential equations predicting the behaviour of the investigated system because it relies on the stability in both space and time of the experimental (observational in the case of astronomy) results. The drastic reduction of the relevant properties to very few basic features like mass and distance, may generally allow for a straightforward prediction and explanation of what is going on and confirm or reject the proposed abstraction. However, a quantitative description in terms of differential equations does not necessarily imply a simple or even a real solution and therefore prediction and explanation could be limited and may require further abstractions. In any case, this was historically the main avenue of "hard sciences" and the reason why a great part of biomathematics redounds around Volterra-Lotka prey-predator models in which both the Godel and Einstein answers to the sponsor have important "real world" consequences.

The framework of Disorganized Complexity (class 2) allows for a still greater generalization power than class 1 by means of a very different style of reasoning. Here, the predictive (and explanatory) power stems from the generation of coarse grain macroscopic descriptors corresponding to gross averages on a transfinite number of atomic elements. Thermodynamics is one of the brightest examples of this statistical approach: emergent collective state variables like temperature or pressure fully describe the system without resorting of full knowledge of microscopic (noise-dominated) details, which can be considered homogeneous.

Class 1 approach asks for few involved elements interacting in a stable way, while class 2 style needs a very large number of identical particles with only negligible (or very stable and invariant) interactions among them. Biological systems only in very few cases do fulfil these constraints, so we step into Weaver's third class (Organized Complexity): the biomedical sciences kingdom.

Organized Complexity arises when many (even if not so many as in class 2) non-identical elements each other interact with time-varying correlation strength. Organized Complexity presents unique features like non-stationarity (this is why some apparently "Disorganized Complexity" situations when out of equilibrium enter the Class 3 domain) and structuring across different mutually interacting organization scales.

The above features generate an extreme context dependence of the results so giving rise to the "information crisis" biology is experiencing [5]. This is the "middle kingdom" where life sciences live that was recognized as the XXI century frontier of basic science [6].

It is worth stressing that the conscious adoption of class 3 style, asks for a deep recasting of both "number crunchers" and "test tube lovers" way of thinking. Quantitatively oriented scientists must accept that contingencies can be more important that general laws. A graph with nodes (indicating players like protein species or genes) linked by edges (empirical correlations between nodes, mutual interactions) is not a proxy of a law of nature but only a specific configuration of the system that does not necessarily happen and could be substituted by an alternative one under different environmental pressures. On the other side, biology-oriented scientists must understand that the pure addition of finer details to an already complicated picture does not generate a most efficient explanation but only increases confusion and irrelevance [7].

As a matter of fact, the relation between "number crunchers" and "test tube lovers", often encompasses a mixing of the first two Weaver's classes: the "number cruncher" offers the "test tube lover" a "statistical significance" obtained by a "rigorous methodology" (Class 2), the "test tube lover" translates these results into a "plausible mechanism" mimicking Class 1 style. Diagrams made of boxes and lines connecting them, very frequent in biological papers and normally referred as "mechanism", for the neat prevalence of contextual information and constraints over general laws, only delineate one out of many possible descriptions of the system at hand. On the other side, the reaching of a "statistical significance" is nothing more than a suggestion of a potentially interesting case and not the seal of "scientific truth": these two epistemological biases concur to the actual reproducibility crisis affecting science [8].

The above sketched liaison between the "biologist" and the "mathematician" does not allow for any fruitful integration between different scientific traditions: the biologist looks at the "mathematical person" as a plumber to call when some hydraulic problem arises, while the "mathematician/plumber" looks condescendingly to whom he/she considers as a self-defining scientist lacking the very bases of quantitative thinking.

Before going ahead, it is worth stressing this is an ultra-simplified (and thus necessarily flawed) picture that does not take into account fields like ecological and environmental studies characterized by a more balanced relation between the two extreme approaches (it is not by chance that emerging fields like the study of microbiome are tailored upon ecological approaches [9]). Nevertheless, this duality is evident in the great majority of the work experiences of the members of our group.

Something changed with the pervasive use of very powerful and cheap computers: biomedical scientists (fascinated by user-friendly sophisticated software) decided that probably they had no more need of that arrogant plumber. They consequently began to use very refined mathematical tools without an adequate knowledge of their applicability and motivations; this in some way made things still worse but prepared the scene for taking seriously the organized complexity character of biological systems.

The deluge of data provoked by high throughput technologies (e.g., omics, neuroimages…) made the "plumber toolbox" anachronistic: the number of variables largely exceeded the number of independent observations putting upside down classical bio-statistical methods. The technological revolution (as often hap-

pens with revolutions) originated an apparently paradoxical phenomenon: the reviving of old traditions by their embedding into a novel perspective. The time-honoured (still alive in protected niches like phytosociology and psychometrics) tradition of multivariate data analysis became essential to get rid of the problems raised by transcriptomic, metabolomic, proteomic and microbiome studies. In the same time, the need of predicting complex outcomes with huge amount of heterogeneous information made machine learning approaches enter into the scene; these computational intensive methods, in order to be explainable (and thus usable in biomedical realms) pushed toward the necessary erosion of the epistemological barrier separating the home-owner and the plumber. Both of them must understand the dynamics of correlation structures (complex networks to use a fashionable term) at the basis of the studied phenomena. The classical separation of scientific enterprises into a linear sequence made of: "hypothesis setting" – "experimental methods" – "data analysis" – "hypothesis verification/falsification" is no more tenable. The mutated conditions introduced new issues spanning the entire research process. Just to name a few: 1) the necessity of standardization of data (along their entire "life cycle") to guarantee their efficient exploitation; 2) the chance of integrating heterogeneous data such as those provided by different "omics" 3) the benefits of appropriate visualization methods for multilevel and multidimensional data to facilitate information extraction and effective communication.

The introduction of machine intelligence went together with the need of "explainability", i.e., the need to conjugate the prediction of a relevant biomedical end-point with a coherent theoretical model validating the obtained result. This urgent need provoked a resurgence of interest in some historical pillars of scientific methodology like Bayesian and dimensionality reduction approaches.

Overall, the new reference frame fostered the resurgence of the quest for integration of different scientific traditions and prompted the collaboration of different kinds of number crunchers (engineers, physicists, statisticians, mathematicians, biophysicists, chemists) and test tube lovers (pharmacologists, biologists, physicians). The entanglement of "content" and "methodological" knowledge is the most promising epistemological novelty made necessary by the actual information crisis and consequent lack of efficacy [5] of scientific research, while the discarding of theory-oriented science in favour of a purely brute-force approach based on an acritical use of informatics tools is a deadly temptation to be avoided [10, 11].

Conflict of interest statement

The Authors declare that there are no conflicts of interest.

REFERENCES

1. Giuliani A, Zbilut JP. The relevance of physical and mathematical modes of thought on complex systems behaviour in biological systems. Complexity. 1998;3(5):23-4.
2. Geman D, Geman S. Opinion: Science in the age of selfies. Proc Natl Acad Sci. 2016;113(34):9384-7. doi: 10.1073/pnas.1609793113
3. Choi BC, Anita WP. Multidisciplinarity, interdisciplinarity, and transdisciplinarity in health research, services, education and policy: 3. Discipline, inter-discipline distance, and selection of discipline. Clin Invest Med. 2008;E41-E48. doi: 10.25011/cim.v31i1.3140
4. Weaver W. Science and complexity. Am Scientist. 1948;36(4):536-9.
5. Ioannidis J. Why most published research findings are false. PLoS Med. 2005;2(8):e124. doi: 10.1371/journal.pmed.0020124
6. Laughlin RB, Pines D, Schmalian J, Stojković BP, Wolynes P. The middle way. Proc Natl Acad Sci. 2000;97(1):32-7.
7. Transtrum MK, Machta BB, Brown KS, Daniels BC, Myers CR, Sethna JP. Perspective: Sloppiness and emergent theories in physics, biology, and beyond. J Chem Phys. 2015;143(1):010901. doi: 10.1063/1.4923066
8. Young SS, Karr A. Deming, data and observational studies: a process out of control and needing fixing. Significance. 2011;8(3):116-20. doi: 10.1111/j.1740-9713.2011.00506.x
9. Gilbert JA, Lynch SV. Community ecology as a framework for human microbiome research. Nat Med. 2019;25(6):884-9. doi: 10.1038/s41591-019-0464-9
10. Bell G, Hey T, Szalay A. Computer science: Beyond the data deluge. Science. 2009;323:1297-8. doi: 10.1126/science.1170411
11. Calude CS, Longo G. The deluge of spurious correlations in big data. Found Sci. 2017;22(3):595-612. doi: 10.1007/s10699-016-9489-4