Organisms
Journal of Biological Sciences

Vol. 3, 1 (July 2019)
ISSN: 2532-5876
DOI: 10.13133/2532-5876_5

INVITED EDITORIAL

Data science is not science
Big data medicine also needs scientific theory

Sui Huang[*]

[*] *Institute for Systems Biology, Seattle WA, USA*

Corresponding author: Sui Huang sui.huang@systemsbiology.org

Citation: Huang, S., 2019 "Data science is not science - Big data medicine also needs scientific theory", *Organisms. Journal of Biological Sciences*, vol. 3, no. 1, p. 1-2. DOI 10.13133/2532-5876_5

One may think that the swings in the history of science between empiricism, which emphasizes observation and experiment ("data") and rationalism, which enjoys knowledge and deductive reasoning ("theory"), have long come to rest with the recognition that both approaches synergize: observations lead to erection of a theory that gives them meaning, and theories must be validated by empirical tests. Theory needs data and data need theory.

But in the era of big data, sadly, deep sequencing, deep learning, and now deep phenotyping, threaten to overtake deep thinking in the life sciences. With the entrance of "data science" (which is not a science) in biomedicine, the dualism of data and theory has re-emerged –but now heavily tilted toward a hardly noticed imbalance: While theorists (who are scientists) espouse, if not crave for data, the data scientists (who are not scientists) eschew, if not contempt theory. Often, they consider the results of data analytics the endpoint of inquiry, and equate them with knowledge.

This asymmetry is exposed in the rushed application of data science to medicine, notably in the new context of "scientific wellness". The recent failure of the scientific wellness start-up Arivale (Senior, 2019), epitomizes the hubris of the approach of "data without theory" (Huang, 2018). What a pity, for the original idea of scientific wellness, championed by Leroy Hood (Schmidt, 2014), is a bold vision and grounded in a well-thought through concept for how the omics technologies could be deployed to improve wellness in a holistic, yet scientific manner that utilizes the new dimensions of personalized data. Unlike in gene-centered precision medicine, one would combine gene sequencing with phenotype profiling of individuals across multiple scales ("deep phenotype") and integrate their personal data with our knowledge of human physiology to promote and stabilize their "wellness" in the N-of-One setting. In doing so, data-driven personal wellness would also overcome the flaws of evidence-based medicine which operates at the level of population averages and ignores individuality (Montori and Guyatt, 2009).

Financial analysts have discussed the end of Arivale in terms of business model and markets (Senior, 2019). But in a more encompassing sphere of thought, Arivale's demise manifests a profound epistemological category mistake: Data science is not science. The start-up's failure to embrace human biology as a science prevented it from successfully implementing Hood's original vision. The expectation that statistical correlations, cluster analysis, trend lines, etc. alone could systematically

inform about medical "actionables" without a thorough understanding of (patho)physiology and the non-linear dynamics that defies many an assumption made in statistical analysis, is folly. In other words, this idea, without scientific theory and without profound domain knowledge beyond the ad hoc looking up in the medical literature and superficial consultation with medical practitioners, is collective naivety at best and innocent hubris at worst.

Whence the hubris? Data-driven approaches have been immensely successful outside of science. The precision with which internet companies predict the movie that you are most likely to see next or the book that you will read next is stunning. But actionable information in these cases can be directly "read off" the results of statistical analysis of your history and your fellow netizen profiles. Similarly, sport data analytics, most vividly presented in the movie "Moneyball", can predict the performance of players with sufficient precision to help clubs make money. No particular theory of stability of athletic performance and no knowledge of sports psychology is needed. The problem is that these use cases do not belong to the category of science or medicine.

In the above cases, the conversion of data patterns, such as a correlation, to useful knowledge, or "actionables", is a purely *transactional* operation, not part of the scientific quest for principles that govern the function of a system, such as an organism. By contrast, medical practice requires an intermediate step between data and the actionables: the science of the organism. Unfortunately, most data scientists are agnostic of the very existence of this step. It includes deductive reasoning and developing theories based on existing scientific knowledge.

Yet, our scientific knowledge of the human body is still far from reliable. Mechanistic understanding of pathophysiology, or of molecular and cellular pathways, while they have in some exceptional cases produced a life-saving game changer -think of Glivec - (Druker, 2008)), is generally not sufficient to offer actionable information for an individual. The science is simply not there yet. It is in fact the spectacular failures of mechanistic rationales that have fostered the rise of evidence-based medicine, which have prepared the climate of thought for data-driven medicine. But as shortcomings of the latter begin to surface, the pendulum of science is swinging back towards appreciation of theoretical principles (Soto et al, 2016). ORGANISMS welcomes this new corrective counter-momentum. Nevertheless, a theory of the organism will have to free itself from the prevailing cult of molecular reductionism, which has side-lined the study of pathophysiology above the level of tissues. Organismal (patho)-physiology is an essential element of a theoretical understanding in medicine - but of course in doing so it must integrate molecular mechanisms as part of the explanation.

With the spread of deep phenotyping, the spate of data could stimulate new theories and permit their validation. It is tempting to engage in data-driven transactions with minimal scholarly consideration of domain-specific theoretical principles. Such approaches can result in actionable knowledge outside of science, but not in medicine where knowledge of biology and physiological principles must accompany data analysis to give meaning to the data. But even then, the deduced potential actionable information must still be tested in clinical trials. You cannot moneyball medicine.

References

Druker, B.J. (2008). Translation of the Philadelphia chromosome into therapy for CML. *Blood* Vol. 112, pag. 4808-4817.

Huang, S. (2018). The Tension Between Big Data and Theory in the "Omics" Era of Biomedical Research. *Perspect Biol Med* Vol. 61, pag. 472-488.

Montori, V.M., and Guyatt, G.H. (2009). Using N-of-1 Trials in Evidence-Based Clinical Practice—Reply. *JAMA Intern Med* Vol. 10, pag. 1022-1023.

Schmidt, C. (2014). Leroy Hood looks forward to P4 medicine: predictive, personalized, preventive, and participatory. *J Natl Cancer Inst* Vol. 106 (12).

Senior, M. (2019). 'Scientific wellness' searches for a business model. *Nature Biotechnology*. (June 12, Commentary)

Soto, AM, Longo, G, Miquel, PA, Montevil, M, Mossio, M, Perret, N, Pocheville, A & Sonnenschein, C 2016. Toward a theory of organisms: Three founding principles in search of a useful integration, *Prog Biophys Mol Biol*, Vol. 122, pag. 77-82.