IJHS

Editorial

Toward self big data

Mohammed Saqr[1], **Sonsoles López-Pernas**[2]

[1]Media Technology and Interaction Design, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden; [2]Departamento de Ingeniería de Sistemas Telemáticos, ETSI Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain

Address for correspondence:
Dr. Mohammed Saqr, Media Technology and Interaction Design,
School of Electrical Engineering and Computer Science,
KTH Royal Institute of Technology, Stockholm, Sweden. E-mail: mmas3@kth.se

WEBSITE: ijhs.org.sa
ISSN: 1658-3639
PUBLISHER: Qassim University

The increasing power of computers, the availability of cheap hardware, and the digitization process overall have made technology an essential part of our daily routines. The massive adoption of technology has helped generate massive amounts of data, often referred to as big data. Such massive data cover all aspects of our life, from dealing with government services (e.g. getting a driving license) to recording sleep patterns with our smartphones.[1] Big data is conceptually more than "lots of data." Big data has several characteristics in addition to huge volumes of data: Rapidly generated in real time or close to, comprehensive, relational, diverse, and fine grained. The concept of big data has been associated with rapidly advancing methods that enable novel ways of analysis that were impossible before. Machine learning, artificial intelligence, data science, and network science are just examples of such methods.[2]

Having access to data, powerful methods of analysis have kindled a wave of optimism that data can solve the world's most unyielding problems. There are, of course, several examples that have stimulated such big hopes. Many companies were able to harness the potentials of big data for their competitive advantage expanding reach and customer base. Several outstanding human innovations give compelling evidence of the power of big data and analytics. Take for example the several online translation services that can translate between several languages without knowing any and with impeccable quality.[1,2]

Such advances have led to the upsurge of data-driven approaches to the degree that many claim that such data-driven methods are a replacement for scientific theory and traditional hypothesis-driven methods. In a much-discussed article, Chris Anderson argues that "we can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.[3]" While in many instances, this has proven useful, in many others, several spurious findings were obtained.[2] Thus, this argument has been largely debated as misleading and dangerous.

A perfect scenario is to use big data to conduct robust analysis with maximized statistical power. Such analyses have been used to develop, for example, product recommender systems and academic dropout prediction systems. However, the data used by such systems are obtained from many individuals in most of the cases; or what is known as variable or group-based analysis. Typically, a large amount of data are collected from a sample that is hypothesized to be representative of the population. The analysis of such sample aims at devising universal laws that apply to all. For instance, the analysis of hypertensive patients would give an idea about the typical presentation of the disease, the typical prognosis, and response to treatment.[4,5] However, such generalization is a mere over-simplification and reductionist.

Group-based analysis has several fundamental problems that pose considerable threat to the validity of results. The first problem is that group-based analysis assumes that the population is a homogenous group of individuals with no or little variations or heterogeneity. Such assumption is very unlikely to hold. Human physiology, pathology, and behavior are heterogeneous with obvious diversity and distinct trajectories. For example, patients with hypertension can have several different pathways of their clinical progression, response to treatment, complications, and disease outcome. Such heterogeneity has been poorly addressed in previous research.[4]

The other problem with group-based analysis is that it assumes that the analyzed process is ergodic, that is, stable overtime, while human social and psychological processes have considerable time variations. Therefore, group-based analysis which lacks the dimension of time can poorly represent these processes. Such deficiencies make results of group-based analysis less likely to generalize to individuals.[4]

The frustration resulting from lack of generalizability of group-based insights is well-recognized. Several threads of research are emerging to address such challenges. Precision medicine, person-based research, and idiography are among the most promising. The main principle behind such methods is to devote more attention to the individual to identify the characteristics that define how their bodies work. Idiography, in particular, aims at collecting data from the individual over multiple times. By having several observations about an individual, researchers have enough data about a person that they can create statistical models based on the person's data. Analysis of single person data offers a laser-sharp view of a person's internal conditions. This approach is made possible by the availability of multiple data sources as well as devices that can facilitate data collection, for example, smartphones.

In summary, a paradigm shift is happening, where the focus is on collecting big data from individuals. Such self-big data is more relevant, more representative, and offers a better opportunity for personalized insights. Such paradigm shift is moving from the big data of many others to self-big data.

References

1. Saqr M. Big data and the emerging ethical challenges. Int J Health Sci (Qassim) 2017;11:1-2.
2. Kitchin R. Big data, new epistemologies and paradigm shifts. Big Data Soc 2014;1:14.
3. Anderson C. The end of theory: The data deluge makes the scientific method obsolete. Wired Mag 2008;16:7-16.
4. Fisher AJ, Medaglia JD, Jeronimus BF. Lack of group-to-individual generalizability is a threat to human subjects research. Proc Natl Acad Sci U S A 2018;115:E6106-15.
5. Saqr M, Lopez-Pernas S. Idiographic Learning Analytics: A single student (N=1) approach using psychological networks. Companion Proceedings of the 11th International Learning Analytics and Knowledge Conference (LAK'21); 2021. p. 456-63.