*Proc Indian Natn Sci Acad* **80** No. 3 September 2014 pp. 677-678


# PREFACE

## Data and Scientific Research

ALOK BHATTACHARYA
*School of Life Sciences, Jawaharlal Nehru University, New Delhi 110 067, India*

Data play an important role in scientific research, either directly related to data generation or data analysis; the later may be based on data generated by other scientists. The importance of data in business decision making as well as in formulating policies for governments have been widely recognized (http://www.ibrc.indiana.edu/studies/Indiana_data_environment.pdf). Unfortunately it has not been made clear how data is needed in scientific discoveries and innovation. One of the definitions of data is "*Information in raw or unorganized form (such as alphabets, numbers, or symbols) that refers to, or represents, conditions, ideas, or objects. Data is limitless and present everywhere in the universe*" (http://www.businessdictionary.com/definition/data.html#ixzz30pUTbrVv). Collation of any information over space or time can be called data. For example, rainfall over different parts of the world (space) and over a period of time (time) can be referred to as a set of data. Collection of such datasets of different parameters can be a valuable tool for scientific discoveries. If we interrogate rainfall data with that of disease prevalence it may be possible for us to hypothesize that increased rain fall may be directly related to higher incidences of infectious diseases and this may help government to come out with policies that will reduce incidences of infectious diseases in areas where rainfall is high. Mining of information and generation of hypothesis by analyzing multiple sets of data is becoming an important area of research and major discoveries are being made using data mining approaches (Fayyad *et al.*, 1996). One of the interesting examples of application of data mining is to provide evidence for an increase in the altitude of malaria distribution in warmer years, which implies that climate change will, without mitigation, result in increase of the malaria burden in the densely populated highlands of Africa and South America (Siraj *et al.*, 2014). This study was carried out using spatiotemporal data at a regional scale in highlands of Colombia and Ethiopia describing the spatial distribution of the disease and variability of temperature. Data mining and machine learning methods have played a key role in diagnosing and predicting flu epidemics (Kofod-Petersen, 2012). Here researchers have harnessed the power of social media to generate data needed for such analysis.

There are a number of issues about data itself. Data formats vary depending upon the nature of data and quite often two datasets do not talk to each other. Therefore, it is important to consider interoperability of data sets while creating new databases. Data standards, formats and errors are important factors and attention has to be paid to develop these and negotiate to obtain one set of international standards that are applicable everywhere. Many datasets are not available for analysis as people who generate the data may not either have plans or do not have the necessary technical skills to carry out analysis. If data can be published and proper system is developed for citation so that people who generate the data get proper recognition, more data will be available for analysis and this will help in enhancing our scientific capabilities. Accessibility of data is a big issue. While complete open access of all data may not always be

*Author for Correspondence: alok.bhattacharya@gmail.com

desirable, it is also not helpful to make the data not accessible at all. Since public support goes in generating much of the scientific data, there is a view that the data generated through public funding should be made accessible with proper credit to the researchers who generated the data. Data related to security issues and those where IPR protection has not been obtained can be excluded. It is also felt that a policy is needed to make use of partial fragmented data that is generated at different places and avoid duplication by generating similar data. There should be a possibility of compilation and curation so that maximum benefit can be achieved from the data. A metadata of all data related work will help to remove some of the problems mentioned above.

Technological advances in data generation and ability to analyze, store and manipulate data at a large scale has helped us to realize immense possibilities in data based research and hypothesis generation. Quantity of data that is being generated in different areas, such as particle physics (Large Hadron Collider), life sciences (genome sequences) and geospatial is at unprecedented scale. The concept of "big data" is permeating all fields of sciences and technology and the challenges involved are considered as major problems to be solved. It will require a concerted effort of large number of experts from different disciplines, such as computer scientists, tele-communication scientists, information scientists, industry, social scientists, legal experts and domain specialists to develop proper approaches, tools, methods, policy framework and mindset to generate, store, distribute and analyze large amount of data. We also need to educate students and scientists about the importance of data and how we can harness this new power to improve our scientific capabilities and enhance innovation. Since data is generated using lot of resources these must be used for societal improvement.

The issues considered here and in the papers in this section will be discussed in detail during the forthcoming conference, SciDataCon-2014, to be held in Jawaharlal Nehru University Convention Centre during November 2-5, 2014. This conference is being jointly organized by Indian National Science Academy, and two International Organizations, CODATA and World Data Centre.

## References

Fayyad U, Piatetsky-Shapiro G and Smyth P (1996) From data mining to knowledge discovery in databases *AI Magazine* **17** 37-54

Kofod-Petersen A (2012) Machine learning and data mining for epidemic surveillance *Medical Jour Australia* **196** 301

Siraj A S, Santos-Vegas, M, Bouma M J, Yadeta D, Ruiz Carrascal D and Pascual M (2014) Altitudinal changes in malaria incidences of Ethiopia and Columbia *Science* **343** 1154-1158.