Chapter 30
Statistical and Computational Needs for Big Data Challenges

Soraya Sedkaoui
Khemis Miliana University, Algeria & SRY Consulting Montpellier, France

ABSTRACT

The traditional way of formatting information from transactional systems to make them available for "statistical processing" does not work in a situation where data is arriving in huge volumes from diverse sources, and where even the formats could be changing. Faced with this volume and diversification, it is essential to develop techniques to make best use of all of these stocks in order to extract the maximum amount of information and knowledge. Traditional analysis methods have been based largely on the assumption that statisticians can work with data within the confines of their own computing environment. But the growth of the amounts of data is changing that paradigm, especially which ride of the progress in computational data analysis. This chapter builds upon sources but also goes further in the examination to answer this question: What needs to be done in this area to deal with big data challenges?

INTRODUCTION

With the advent of digital technology and smart devices, a large amount of digital data is being generated every day. Individuals are putting more and more publicly available data on the web. Many companies collect information on their clients and their respective behavior. As such, many industrial and commercial processes are being controlled by computers. The results of medical tests are also being retained for analysis. Financial institutions, companies, and health service providers, administrations generate large quantities of data through their interactions with suppliers, patients, customers, and employees. Beyond those interactions, large volumes of data are created through Internet searches, social networks, GPS systems, and stock market transactions.

This brings us to think about the legend of the wise 'Sissa' in India. When King 'Belkib' asked about the reward he desired, after his invention, he asked to receive a grain of rice for the first square, two grains for the second, four grains for the third and so on. The king agreed, but he didn't know that on the last square of the board he should drop 2^{63} grains, or more than 700 billion tons. In their book

DOI: 10.4018/978-1-6684-3662-2.ch030

"Race Against the Machine," Brynjolfsson and Mcaffee (2011) referenced the fable of the chess and rice grains to make the point that "exponential increases initially look a lot like linear, but they are not. As time goes by – as the world move into the second half of the chessboard – exponential growth confounds our intuition and expectation".

Thus currently, not only is the quantity of digitally stored data much larger, but the type of data is also very varied, thanks to the various new technologies (Sedkaoui & Monino, 2016). Data volume will continue to grow and in a very real way, the data produced, as well as other data accumulated, constitutes a constant source of knowledge. This widespread production of data has resulted in the 'data revolution' or the age of 'big data'. Big data gets global attention and can be best described using the three Vs: volume, variety and velocity. These three dimensions often are employed to describe the phenomenon. Each dimension presents both challenges for data management and opportunities to advance decision-making. In another way, every data tells a story and data analytics, in particular the statistical methods coupled with the development of IT tools, piece together that story's reveal the underlying message.

This 3 V's provide a challenge associated with working with big data. The volume put the accent on the storage, memory and computes capacity of a computing system and requires access to a computing cloud. Velocity stresses the rate at which data can be absorbed and meaningful answers produced. The variety makes it difficult to develop algorithms and tools that can address that large variety of input data. So, there are still many difficulties and challenges in the use of big data technologies. And, if decision-makers can't understand the power of data processing and analytics, they may be, in some ways, the "Belkibs" of big data value. The key is applying proper analytics and statistics methods to the data. Thus, from this data companies derive information and then producing knowledge, or which it called the target paradigm of "knowledge discovery", described as a "knowledge pyramid" where data lays at the base. To advance successfully the paradigm effectiveness data analysis is needed.

The analysis of big data involves multiple distinct phases which include data acquisition and recording, information extraction and cleaning, data integration, aggregation and representation, query processing, data modeling and analysis and interpretation. These all are the methods of modern statistical analysis necessary for dealing with big data challenges. But, each of these phases introduces other challenges: Heterogeneity, scale, timeliness, complexity, quality, security...

Modern data analysis is very different from other methods which existed prior. Also, data is very different from data which existed before. In another word, the nature of modern data (greatest dimension, diverse types, mass of data) does not authorize the use of most conventional statistical methods (tests, regression, classification). Indeed, these methods are not adapted to these specific conditions of application and in particular suffer from the scourge of dimension. These issues should be seriously considered in big data analytics and in the development of statistical procedures.

Consider a simple example to explain a quantitative variable Y through a set $\{X1, \ldots Xp\}$ of quantitative variables: $Y = f(X1, \ldots Xp) + \varepsilon, [(yi, xi), i = 1,. . ., n]$

If the function is assumed to be linear and p is small, on the order of ten; the problem is well known and widely discussed in the literature. In the case where the function f is not exactly linear and n is large, it is possible to accurately estimate a larger number of parameters and therefore to envisage more sophisticated models. Keeping to the usual Gaussian model, even the simplest case of a polynomial model quickly becomes problematic. Indeed, when the function is linear, take $p = 10$, the model selection procedure is facing a group of 2^{10} possible models and shrewd algorithms allow to cope.

However, consider to estimate f, a simple polynomial of second or third degree, with all its interactions, leads us to consider a large number of parameters and thus, by combinatorial explosion, an astronomical