Chapter 4
Defining, Understanding, and Addressing Big Data

Trevor J. Bihl
Air Force Institute of Technology, USA

William A. Young II
Ohio University, USA

Gary R. Weckman
Ohio University, USA

ABSTRACT

"Big Data" is an emerging term used with business, engineering, and other domains. Although Big Data is a popular term used today, it is not a new concept. However, the means in which data can be collected is more readily available than ever, which makes Big Data more relevant than ever because it can be used to improve decisions and insights within the domains it is used. The term Big Data can be loosely defined as data that is too large for traditional analysis methods and techniques. In this article, varieties of prominent but loose definitions for Big Data are shared. In addition, a comprehensive overview of issues related to Big Data is summarized. For example, this paper examines the forms, locations, methods of analyzing and exploiting Big Data, and current research on Big Data. Big Data also concerns a myriad of tangential issues, from privacy to analysis methods that will also be overviewed. Best practices will further be considered. Additionally, the epistemology of Big Data and its history will be examined, as well as technical and societal problems existing with Big Data.

1. INTRODUCTION

Although various definitions exist for Big Data, as noted by Marcus (2013), few know precisely what the term means. Big Data is commonly defined with vague and self-referencing definitions. Frequently these definitions include buzzwords, which cause Big Data to be viewed as a "buzzword" itself (Jedras, 2013; Global Language Monitor, 2013). However, the authors do not share this view and will quickly dispel that notion by examining the phenomenon of Big Data, reviewing disciplines that consider it an

DOI: 10.4018/978-1-5225-7501-6.ch004

issue, exploring various problems both associated with Big Data and caused by Big Data, and examining methods of handling Big Data and its problems. One aspect driving Big Data is that data storage capabilities that continuously grow and hence allow more data to be stored. This is captured in a revision of Parkinson's Law (1961), "data expands to fill the space available for storage," as noted by (Kleinstein, 2005; Sarafis, Trinder, & Zalzala, 2007).

First, the authors will explore defining Big Data, which involves examining the historicity of data, and addresses the differences between data and Big Data. Such an analysis leads directly to understanding Big Data and its impact in various domains. A few observations will be evident, including that the issue of Big Data is not new, but that its appearance in a multitude of domains is. Due to various issues related to Big Data, addressing and handling Big Data are critical and a major focus herein. As a variety of contemporary issues exist in Big Data; see (Gudivada, Baeza-Yates, & Raghavan, 2015) these should be examined. The authors further consider the various computational and methodological means to process Big Data.

Tangential concepts are also explored. Despite some successes in exploiting Big Data, many companies either do not have a plan to exploit Big Data or do not know how to go about it (Bartram, 2013). While Big Data is considered revolutionary, it is arguable that small datasets and feature selection can result in more focused analysis (Arbesman, 2013; Bartram, 2013, p. 30); therefore, this matter needs to be considered. The authors posit that *value* or *saliency* of Big Data is *a*, if not *the*, critical component in exploiting Big Data, and thus the various means to process Big Data are a significant focus herein.

This paper is organized as follows: a general understanding of what is data, Big Data, and its history is presented next in Section 2. Section 3 then discusses various issues with Big Data and methods to address Big Data. Section 4 then concludes the paper.

2. UNDERPINNINGS OF BIG DATA

To understand Big Data, one must first understand data. Beyond this simple distinction, and to vet the various expectations associated with Big Data, one must further understand how data is related to information, knowledge, and wisdom.

2.1. What Is Data?

Data itself can consist of all forms of sensing outputs, data logging outputs, and collected information. In general, data is contained and shared by four storage mediums: print, film, magnetic, and optical (Lyman P., et al., 2003). Thus, data can include remotely sensed images taken from an airplane, archival pdf documents stored on a hard drive, books in a library, patient medical records, etc. However, discerning relevant information, knowledge, wisdom, or actionable information from raw data is another matter (Stenmark, 2002).

Data, information, knowledge, and wisdom (DIKW) are higher-level abstractions that are inter-related; Figure 1 in the Appendix conceptualizes their relationship in the continuum. Although inferring a linear relationship between data-information-knowledge is not considered accurate (Stenmark, 2002), Figure 1 does illustrate that a hierarchy exists between DIKW. Furthermore, the definition and distinction between the quantities of DIKW are metaphysical (i.e. not physically measurable (Rickaby, 1890) in nature) sub-